Bayesian inference and reinforcement learning models in decision making tasks

Universidad Autónoma de Madrid

Departamento Física Teórica

Facultad de Ciencias

Stefania Sarno

Memoria de Tesis presentada para optar al grado de

Doctor en Biofísica

Director: Néstor Parga Carballeda

September 2017

Abstract

Every day animals and humans must make countless decisions that lead to better long-term outcomes in order to survive. Learning to associate unambiguous sensory cues with rewarded choices is known to be mediated by dopamine neurons. From a computational standpoint, reinforcement learning constitutes the normative framework to learn optimal behaviour, and it has been fundamental in understanding how reward-related brain areas works in simple paradigms of conditioning. In particular, convincing evidence, first inspired by algorithmic ideas from reinforcement, seems to indicate that the activity of dopamine neurons signals reward prediction errors, or the discrepancy between the expected and currently experienced rewards. However, little is known about how dopamine neurons behave when choices rely on such uncertain reward-predicting stimuli.

In this thesis I analyse whether and how dopamine neurons represent reward prediction-related signal in tasks that require non-trivial processing of external information. When sensory information is ambiguous and incomplete it has been suggested that the brain engages in Bayesian inference during perception. Assuming that reward predictions and reward prediction errors are computed using the result of this inference process we propose a model based analysis of the dopamine signal during two different tasks.

In chapter 3 I re-examine data recorded when the animal is engaged in the detection of possibly weak vibrotactile stimuli delivered at random times. According to this study the dopamine reward prediction error signal reflects the certainty of the animal about the detection and temporal expectations that depend on the subjective detection of the relevant stimulus.

In chapter 4 I analyse new recordings collected during a discrimination task requiring the sequential comparison of two vibrational stimuli separated by a delay period of a few seconds. I find that the DA reward prediction error signal in response to the second stimulus codes for the subjective difficult that the animal faces in the discrimination process. According to our analysis this subjective difficult is related with differential sensory evidence integration (that differ from trial-to-trial) and the dopamine activation is in line

with an anomaly in the performance, known as the contraction bias that can emerges from Bayesian inference. Additionally we find a sustained positive tonic activation of the dopamine neurons during the delay period. This persistent activation has not been encountered a previous studies involving working memory, but is consistent with the idea that tonic dopamine plays an important role to retain the relevant information in working memory.

The results of this thesis extend the well known reward prediction error coding of dopamine to the domain of decision making, and suggest a normative framework to study the role of these neurons in guiding optimal behaviour.

Resumen

Todos los días, para poder sobrevivir, tanto animales como humanos deben tomar decisiones que lleven a buenos resultados a largo plazo. Es sabido que las neuronas dopamina median en el proceso de aprender a relacionar estímulos concretos con decisiones que lleven a una recompensa. Desde un punto de vista computacional, el aprendizaje con refuerzo constituye el marco normativo para aprender comportamientos óptimos, y ha sido fundamental para entender como áreas del cerebro relacionadas con recompensa funcionan en casos simples de condicionamiento. Particularmente, hay evidencia convincente, inspirada inicialmente por ideas de algoritmos de aprendizaje con refuerzo, de que la actividad de las neuronas dopaminérgicas codifica el error de predicción de la recompensa, es decir, la diferencia entre la recompensa esperada y la recibida. Sin embargo, se sabe poco acerca de como se comportan las neuronas dopaminérgicas cuando las decisiones se basan en estímulos ambiguos.

En esta tesis analizo si, y de que manera, las neuronas dopaminérgicas representan el error de predicción de la recompensa en tareas que requieren procesamiento no trivial de señales externas. Cuando la información sensorial es ambigua e incompleta, se ha sugerido que el cerebro lleva a cabo inferencia bayesiana durante la percepción. Asumiendo que las predicciones de recompensas futuras y los errores de predicción en la recompensa se calculan basándose en esta inferencia, proponemos un modelo basado en el análisis de las señales dopamina en dos tareas distintas.

En el capitulo 3, re-examino datos tomados mientras el animal está involucrado en la detección de posibles estímulos débiles, administrados en tiempos aleatorios. Según este estudio, la señal de error de predicción de la recompensa de la dopamina refleja la certidumbre del animal sobre la detección, así como expectaciones temporales que dependen de la detección subjetiva del estímulo relevante.

En el capitulo 4, analizo nuevos datos tomados durante una tarea de discriminación que consiste en comparar dos estímulos vibracionales aplicados con una diferencia temporal de unos pocos segundos. Encuentro que la señal de error de predicción de la recompensa

de la dopamina después del segundo estimulo codifica la dificultad subjetiva a la que el animal se enfrenta en la tarea. Según mi análisis, esta dificultad subjetiva está relacionada con la integración diferencial de la evidencia sensorial (que difiere de ensayo a ensayo); y la activación de la dopamina está en concordancia con una anomalía que se observa en el rendimiento del animal, conocido como el prejuicio de contracción (*contraction bias*), que puede aparecer fruto de la integración bayesiana. Asimismo, encontramos una activación sostenida de las neuronas dopamina durante el periodo entre estimulos. Esta activación persistente no se había encontrado en otros estudios relacionados con la memoria de trabajo, pero es consistente con la idea de que la actividad de neuronas dopamina juega un papel importante en la retención de información importante en la memoria de trabajo.

Los resultados de esta tesis extienden el conocido rol de la neuronas de dopamina en la codificación del error de la predicción de recompensa al terreno de la toma de decisiones, y sugieren un cuadro normativo para estudiar el rol de estas neuronas en guiar un comportamiento óptimo.

Contents

Chapter 1

Introduction

In this thesis I try to elucidate how dopamine neurons behave when choices rely on uncertain reward-predicting stimuli that require non-trivial processing of external information. Extensive literature exists on the role of dopamine neurons in guiding reward predictions and associative learning in simple experiment of conditioning. From a theoretical standpoint the computational field of reinforcement learning constitutes the normative framework to analyse an interpret the responses of dopamine neurons. During the last years indeed, algorithmic ideas from reinforcement learning clarified the crucial role of dopamine neurons in broadcasting a fundamental teaching signal and suggested possible underlying neural substrates for learning and action selection in the specific brain areas.

This chapter aims to provide an overview of relevant work in three broad fields: computational reinforcement learning, relevant data concerning the activity of dopamine-producing neurons (preceded by a brief introduction to the behavioural paradigm of conditioning and a quick glance to reward related brain areas), and models of the dopamine system.

1.1 Computational Reinforcement Learning

The idea that we learn by interacting with our environment is probably the first that comes up to our mind when we think about the notion of learning. Reinforcement learning, an area of machine learning inspired by behaviourist psychology, constitutes the computational normative framework to deal with the idea of learning from interaction. Generally speaking the reinforcement learning problem is characterized from three distinguishing aspects: first of all, it concerns closed-loop problems, because the learning system's actions influence its later inputs. Second it relies on the so call trial-and-error search: the

learner is not told which actions to take and must be able to collect information about the cause/effect structure of the external environment from its own experience in order to choose the actions leading to the most favorable consequences. Finally, because rewards are often delayed, the learner must face a problem known as the credit assignment problem, i.e the difficulty of relating actions with their long term consequences.

This section will review some basic ideas, notations and algorithms of computational reinforcement learning, mainly focusing on the theoretical work that has been widely applied to the understanding of the brain.

1.1.1 The agent environment interaction

Reinforcement learning concerns the problem of optimal decision making: how we learn to act in an uncertain and changing world in a way that maximizes some definition of reward on the long term. The learner and decision-maker, typically referred to as the agent, is not told which actions to take (as in many forms of machine learning) but has to acquire the ability to select appropriate actions exclusively via the interaction with the environment.

Considering a discrete temporal evolution the agent-environment interaction can be sketched as follows: at each time step t the agent receives information about the state of the environment s_t performs an action a_t, observes the outcome of its actions r_{t+1} and transitions to a (possibly) new state s_{t+1}. The formal definition of the problem requires to specify a set of states $\mathcal{S}$ (representing the situations the agent comes across while interacting with the environment), a set of actions $\mathcal{A}$ (which can differ from state to state, in which case they are denoted $\mathcal{A}(s)$), and two objects describing the dynamics of the environment: a state transition matrix $\mathcal{T}$ whose elements represent the transition probability from all states to their successors, and a vector reward function $\mathcal{R}$ representing the expected value of the immediate reward from all states. Hereafter the shorthand $\mathcal{T}^a_{ss'} = P(s_{t+1} = s'|s_t = s, a_t = a)$ will be used for the elements of the transition probability matrix and $\mathcal{R}^a_s = E(r_{t+1}|s_t = s, a_t = a)$ for the elements of the expected reward function. The fact that the transition probabilities and reward expectations can be specified this way stems from the assumption that the environment is *Markovian*. This means that all relevant information about the system at time t is compactly retained in the last state s_t and action a_t[1].

[1]In the general case given the history of the system $H_t = s_0, a_0, r_0, \ldots, a_t, s_t$ (i.e all the past states, actions and rewards) the elements of the matrix $\mathcal{T}$ are defined as $P(s_{t+1}|H_t)$ and the elements of $\mathcal{R}$ as $P(r_{t+1}|H_t)$, i.e transition and reward probabilities are conditioned on entire history.

While interacting with the environment the agent selects actions according to a policy π: a mapping from states to actions. The mapping need not to be deterministic, so in general a policy represents a mapping from states to probability of actions. The function $\pi(s,a) = P(a|s,\pi)$ indicates the probability of taking the action a given the state s and following the policy π. In this framework the aim of the decision process is dual: first the agent needs to learn which of the visited states are more responsible for future rewards according to its current policy $\bar{\pi}$. This problem is called *policy evaluation* and relates to the well known the *credit-assignment problem*: rewards, especially in fine grained state-action spaces, can occur terribly temporally delayed and therefore the agent needs to take into account not only the immediate consequences of an action (i.e. the immediate reinforcement), but also the value of the next state in term of long-term return. Second the agent needs to find an optimal policy π^*, i.e a mapping from states to action that maximizes future rewards. This process is called *policy improvement*.

A final remark before proceeding with the description of the methods for policy evaluation and policy improvement. In the first part of this chapter I will restrict the attention to the case of "fully observable" environment. This corresponds to assume that the information received from the agent precisely represents the state of the environment. When the environment is both Markovian and fully observable the decision process is called a Markov decision process (MDP). However in real world situations the agent receives only noisy and uncertain informations and is not aware of the precise state of the environment. In this case the environment is only "partially observable" and the decision process is called a partially observable Markov decision process (POMDP). This last approach represents a more realistic mathematical framework for modeling decision making and it will be discussed at the end of this section (in subsection 1.1.5).

1.1.2 Temporal-difference methods for prediction learning

Prediction learning can be studied considering an MDP in which the policy is fixed. In this case the MDP reduces to a *Markov chain* with the transition probability defined as $\mathcal{T}^{\pi}_{ss'} = \sum_{a\in\mathcal{A}(s)} \pi(s,a)\mathcal{T}^{a}_{ss'}$ and reward expectation as $\mathcal{R}^{\pi}_{s} = \sum_{a\in\mathcal{A}(s)} \pi(s,a)\mathcal{R}^{a}_{s}$.[2] Prediction learning relies on the concept of *state-value function* $\tilde{V}^{\pi}(s)$, representing some measure of the expected future reward, known as *expected return*, when the agent starts in state s at time t and follows the policy π. The definition of the state-value function requires

[2]More generally an MDP reduces to a Markov chain (or Markov reward process) whenever the dynamics of environment is independent from the actions of the agent, and transition probability and reward expectation reduce respectively to $\mathcal{T}_{ss'} = P(s_{t+1} = s'|s_t = s)$ and to $\mathcal{R}_s = E(r_{t+1}|s_t = s)$.

to specify the objective of the learning process in a formal way, i.e. we need to define a specific function of the rewards sequence representing the expected return (that is object to predict in the case of prediction learning and the object to maximize when we consider the more general problem of finding optimal behaviour). In *episodic tasks*[3], i.e. those in which the agent-environment interaction breaks naturally into subsequences (usually referred as episodes), it is possible to simply define the long-run return as the cumulative expected future reward.

However, in general, when the temporal horizon of the problem we are interested in is infinite (i.e we study a problem without an obvious endpoint) such a quantity is not finite and a better measure of optimal behaviour needs to be specified. A way to define the finite return in an infinite horizon problem is by introducing the additional concept of temporal discounting. According to this approach, the agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. Usually the long-run return G_t at time t is defined as the exponentially discounted sum of future rewards:

$$G_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \tag{1.1}$$

where γ is a parameter called the discount factor and $0 < \gamma < 1$. At behavioural level discounting allows to explain humans and animals preference for earlier rewards to later ones. The existence of a discount reflects the fact that delaying a reward introduces additional risks because rewards could not be available anymore in the future. The particular form of discounting introduced in Equation 1.1 is known as *exponential discounting*. It is equivalent to assume a constant probability of risk per unit time of a hazard (preventing the reward availability) to occur, given that it has not occurred yet. In what follows Equation 1.1 will be used as the definition for the long-run return.

A few alternative definitions have been use in the RL framework to study infinite horizon problems and will be discussed at the end of this section. Given the definition in Equation 1.1 the state-value function $\tilde{V}^{\pi}(s)$, i.e the expected return when the agent

[3]An absorbing Markov process is a Markov chain in which every state can reach a final state, known as absorbing state, that, once entered, cannot be left.

starts in state s at time t and follows the policy π is defined as:

$$\begin{aligned}\tilde{V}^{\pi}(s) &= E^{\pi}\Big[G_t|s_t=s\Big] = E^{\pi}\Big[r_{t+1}+\gamma\sum_{i=0}^{\infty}\gamma^i r_{t+i+2}|s_t=s\Big] \\ &= \sum_{a\in\mathcal{A}}\pi(s,a)\Big[\mathcal{R}_s^a+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}_{ss'}^a\tilde{V}^{\pi}(s')\Big] \\ &= E^{\pi}\Big[r_{t+1}+\gamma\tilde{V}^{\pi}(s_{t+1})|s_t=s\Big] \\ &= \mathcal{R}_s^{\pi}+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}_{ss'}^{\pi}\tilde{V}^{\pi}(s') \qquad (1.2)\end{aligned}$$

Equation 1.2 is known as the *Bellman equation* for the state-value function under the policy π. It expresses a recursive relationship between consecutive states and can be expressed concisely using matrices:

$$\tilde{\mathbf{V}}^{\pi} = \mathcal{R}^{\pi}+\gamma\mathcal{T}^{\pi}\tilde{\mathbf{V}}^{\pi} \qquad (1.3)$$

where $\tilde{\mathbf{V}}^{\pi}$ is a column vector with one entry per state representing the corresponding state-value. Considering a N-dimensional state space $\mathcal{S}$ once the dynamics of the environment is known (i.e the transition and reward probabilities) the Bellman equation reduces to nothing but a set of N linear equations with N unknowns and can easily be solved[4]. When the number of states is large the problem becomes computationally too complex and difficult to handle. In this case the Bellman equation can be solved with iterative methods.The function $\mathbf{V}^{\pi}$ indicates the N-component vector representing an estimate of the vector $\tilde{\mathbf{V}}^{\pi}$. After initializing each component of $\mathbf{V}^{\pi}$ to zero, the estimate can be repeatedly updated as:

$$\mathbf{V}_n^{\pi} \leftarrow \mathcal{R}^{\pi}+\gamma\mathcal{T}^{\pi}\mathbf{V}_{n-1}^{\pi} \qquad (1.4)$$

where $\mathbf{V}_n^{\pi}$ indicates the nth update of the estimate $\mathbf{V}$ of the real state-value vector $\tilde{\mathbf{V}}^{\pi}$. This approach is guaranteed to converge because it directly implements the consistency relation of Equation 1.3[5]. Solving the Bellman equation for a known MDP represents the core of *Dynamic Programming*. However when the MDP is unknown (i.e the model of the environment is not available) the agent needs to somehow estimate the value function

[4]Equation 1.3 can be written as $\tilde{\mathbf{V}}^{\pi} = (\mathcal{I}-\gamma\mathcal{T}^{\pi})^{-1}\mathcal{R}^{\pi}$, where $\mathcal{I}$ indicates the identity matrix. Therefore the problem reduces to the inversion of the matrix $(\mathcal{I}-\gamma\mathcal{T})^{-1}$, a problem of computational complexity $O(N^3)$ for N states.

[5]This reasoning works in the synchronous form of value iteration, in which the values of all the states are updated simultaneously (i.e. the nth value of every state is used to compute the n + 1st backup of the value function). It turns out an asynchronous form of value iteration also converges. In this version, the value of each state is updated individually using the latest available values of the other states.

from its stream of experience. Temporal difference (TD) methods provide an efficient solution to the problem of prediction learning in a model-free fashion (i.e for an unknown MDP). The key idea behind the TD algorithm is using a guess of the expected long-run return when the agent starts in state s and follows the policy π to estimate the state-value function $V^\pi(s)$. Suppose that the agent starts from state s, performs action a, transitions to state $\bar{s}$ and obtains the outcome r. In this transition the agent experiences a (biased) sample of $\left[\mathcal{R}_s^\pi + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}_{ss'}^\pi \tilde{V}^\pi(s')\right]$. After many visits to state s the agent obtains unbiased samples of the reward expectation and of the transition probabilities [6]. Therefore after each visit to the state s the agent can use the sample it observes (that is $r_{t+1} + \gamma V(s_{t+1})$ as a guess for the value of $\tilde{V}^\pi(s)$ and can update the estimated value of s proportionally to is the discrepancy between what was expected (the current estimate $V^\pi(s_t)$) and what actually occurred (the actual guess). The last quantity is known as the reward-prediction error or TD-error δ_t:

$$\delta_t = [\tilde{G}^{(1)}(s_t) - V^\pi(s_t)] = [r_{t+1} + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)] \tag{1.5}$$

where $\tilde{G}^{(1)}(s_t) = r_{t+1} + \gamma V^\pi(s_{t+1})$ defines the 1-step guess or 1-step return. The dependence on π has been omitted in the 1-step guess for simplifying the notations. The same convention will be adopted for the more general n-step return and λ-return defined below. TD learning then updates the estimate for the value of the last visited state s_t based on the TD error and a parameter α representing the learning rate:

$$V(s_t) = V(s_t) + \alpha \delta_t \tag{1.6}$$

The above equation is known as the $TD(0)$ algorithm or the 1-step TD backup for the state-value function. It is interesting to note that according to Equation 1.6 the state-value function converges to an exponential recency weighted average of the sequence of guesses about the value $\tilde{V}^\pi(s)$ that the agent experiences each time it visits the state s.[7]

The formal justification for TD learning as a method for optimal learning stems from the fact that it directly implements dynamic programming methods (Barto et al., 1989; Sutton, 1988). Under some technical assumptions about the learning rate schedule and the structure of the MDP (e.g., all states must be sampled infinitely often) this algorithm

[6] The idea is that each time the agent visits the state s it selects a with probability $\pi(s, a)$. Therefore after visiting s many times the agent can obtain an unbiased sample of $\mathcal{T}_{ss'}^\pi$ (because after the action a from the state s the environment transitions to the state s' with probability $\mathcal{T}_{ss'}^a$). A similar reasoning applies to $\mathcal{R}_s^\pi$.

[7] Given a sequence of number $x_1, x_2, \ldots, x_n$ and the update $\bar{x}_n = \bar{x}_{n-1} + \alpha[x_n - \bar{x}_{n-1}]$ the value $\bar{x}_n$ converges to $\bar{x}_n = \alpha \sum_{k=0}^{n-1} (1-\alpha)^{n-k} x_k$, i.e an exponential recency weighted average of the sequence.

converges (Dayan, 1992). Arbitrary function approximation schemes can be used to represent the value function V in place of lookup tables and the algorithm has been proven to converge (to some approximation of V whose error can be quantified) for linear function approximation (Bertsekas and Tsitsiklis, 1996). Temporal difference learning with function approximation will be discussed in subsection 1.1.4.

Alternatively the value of the state s can be learned using a n-steps TD backup: after visiting the state s at time t the agent can wait n steps, observe the sequence of rewards $r_{t+1}, r_{t+2}, \ldots, r_{t+n+1}$. Then this sequence of rewards can be used to construct the guess $\tilde{G}^{(n)}(s_t) = r_{t+1} + r_{t+2} + \cdots + r_{t+n+1} + \gamma^n V(s_{t+n+1})$ and to update the value of s_t as

$$V(s_t) = V(s_t) + \alpha[\tilde{G}^{(n)}(s_t) - V(s_t)] \tag{1.7}$$

The validity of the above backup relies on the fact that the Bellman equation can easily be generalized to an equation relating the values of two states separated by n time-steps:

$$\tilde{\mathbf{V}}^\pi = \mathcal{R}^\pi + \gamma\mathcal{T}^\pi\mathcal{R}^\pi + \gamma^2(\mathcal{T}^\pi)^2\mathcal{R}^\pi + \cdots + \gamma^n(\mathcal{T}^\pi)^n\mathcal{R}^\pi + \gamma^n(\mathcal{T}^\pi)^n\tilde{\mathbf{V}}^\pi \tag{1.8}$$

In the more general case every combination of the n-steps guesses $\tilde{G}^{(n)}(s_t)$ can be used to update the value of the state s_t visited at time t. In the TD(λ) algorithm the actual guess for the value function is constructed as the average of all the n-step backups, each weighted proportional to λ^{n-1} , where $\lambda \in [0, 1]$, and normalized by a factor of $1 - \lambda$ to ensure that the weights sum to 1. The λ-guess can be written as

$$\tilde{G}^{(\lambda)}(s_t) = (1 - \lambda)\sum_{i=1}^{n} \lambda^{i-1}\tilde{G}^{(i)}(s_t) \tag{1.9}$$

The value of s_t is updated as:

$$V(s_t) = V(s_t) + \alpha[\tilde{G}^{(\lambda)}(s_t) - V(s_t)] \tag{1.10}$$

The backup of the above equation is not directly implementable online because it is acausal, it requires knowledge of what will happen many steps later the actual time t. However it is possible to (approximately) implement the backup of Equation 1.10 in an online fashion. This can be achieved by introducing an additional memory variable associated with each state, its eligibility trace (Sutton and Barto, 1998), that record which states have been visited recently (in terms of $\gamma\lambda$). Two types of eligibility traces that slightly differ for their temporal evolution have been traditionally used in the RL literature. In the first type, known as *accumulating traces*, the eligibility traces evolve in

time as follow: on each time step the traces decay by a factor $\gamma\lambda$. In addition, the trace of the visited state is increased by 1:

$$e_t(s) = \begin{cases} \gamma\lambda e_t(s) + 1 & \text{if} \quad s_t = s \\ \gamma\lambda e_t(s) & \text{otherwise} \end{cases} \tag{1.11}$$

where the function denoted $e_t(s) \in \mathbb{R}^+$ represents the eligibility trace for the state s at time t. The latter type, known as *replacing traces*, differs from the previous case in the temporal evolution of the trace of the last visited state, which simply takes the value 1 after the visit. The temporal evolution in the replacing case can be written as:

$$e_t(s) = \begin{cases} 1 & \text{if} \quad s_t = s \\ \gamma\lambda e_t(s) & \text{otherwise} \end{cases} \tag{1.12}$$

The TD(λ) update for the value functions at time t is in both cases as follows:

$$V(s) = V(s) + \alpha\delta_t e_t(s) \quad \forall s \in \mathcal{S} \tag{1.13}$$

Therefore, unlike the TD(0) backup Equation 1.13, according to the TD(λ) backup on each time step the value of all the states is updated. Changes are proportional to the reward-prediction error δ_t and to the eligibility traces, that indicate the degree to which each state is responsible for the actual reinforcement. Said in other words: earlier states are given less credit for the actual TD error δ_t and therefore are modiefied less. In this sense eligibility traces provide an elegant solution to the credit-assignment problem. A final remark about eligibility traces. The theoretical view of eligibility traces in term of the λ-return is known as *forward view*, the mechanistic implementation of the algorithm in Equation 1.10 in term of the function $e(s)$ is known as *backward view*. It can be shown (Sutton and Barto, 1998) that, for episodic tasks, and provided that updates are accumulated within episode but applied in batch at the end of episode (*offline* updates) the forward and the backward view of the algorithm are equivalent[8]. When updates are applied *online*, i.e at each step within episode, the forward and backward views of the TD(λ) algorithm are not exactly equivalent. However it has been recently proved that, by using a slightly different form of eligibility trace, a perfect equivalence can be achieved for the online backup (van Seijen and Sutton, 2014).

All the TD(λ) methods presented so far rely on the hypothesis of exponential discounting and on the resulting definition for the long-run return described in (1.1). There exist

[8]Here equivalence indicates that the total update at end of an episode is the same. To be more precise the equivalence holds only when considering accumulating traces.

alternative definitions for the discounting and the rest of this section will quick glance over them.

Many studies show that animals exhibit time preferences which are not exponential, but instead fall off with delay at a decreasing proportional rate. This trend is consistent with *hyperbolic discounting*. Under the assumption of hyperbolic discounting it is possible to define a *hyperbolically discounted return* G_t^H as:

$$G_t^H = \sum_{i=0}^{\infty} r_{t+i+1}/(i+1) \tag{1.14}$$

Hyperbolic discounting can be formally explained by an uncertain underlying hazard rate, with an exponential prior distribution for the hazard to occur (Sozou, 1998). The above definition of the long-term return has been mostly ignored in reinforcement learning literature because the resulting state value function cannot be written recursively and therefore cannot be calculated by recursive methods (such as TD learning; Daw and Touretzky, 2000). Although reinforcement learning methods with hyperbolic discounting will be not discussed in this thesis it is interesting to point out that a recursive temporal difference implementation of the hyperbolic model has been recently proposed in (Alexander and Brown, 2010).

The formulation of temporal decision making in terms of discounting correspond to assume that humans and animals try to maximizes the reward-to-risk ratio. An alternative hypothesis is that animals seek to maximize their average intake of rewards over time. In this case the objective of learning becomes the *average reward return* G_t^A and can be defined as:

$$G_t^A = \sum_{i=0}^{\infty} (r_{t+i+1} - \rho) \tag{1.15}$$

where $\rho = \lim_{n\to\infty} \frac{1}{n} E(\sum_{i=0^n} r_{t+1+i})$ represents the average reward per time-step[9]. A precise TD algorithm for average return reward can be find in (Tsitsiklis and Van Roy, 1999, 2002) and its application to the study of the dopamine neurons response has been explored in (Daw and Touretzky, 2000; Daw et al., 2006).

[9]More precisely given a start state s_t and a stationary policy π the average reward per time-step is defined as: $\rho^\pi = \lim_{n\to\infty} \frac{1}{n} E_{s_t}(\sum_{i=0^n} r_{t+1+i} | s_t, \pi)$. If the MDP is *ergodic* (each state is visited an infinite number of times and without any systematic period) for any stationary policy the average reward per time-step ρ^π that is independent of start state s_t. The same property is valid for an ergodic Markov chain. In Equation 1.15 we are making the implicit assumption of ergodicity.

1.1.3 The Actor/Critic method

This section describes policy improvement algorithms mainly focusing on a method for action selection known as Actor/Critic model. the discussion will focus on the Actor/Critic method because it is the RL method that has been most strongly linked to prediction learning and action selection in the brain, and because it constitutes the core implementation of the models presented in the rest of this thesis (in chapter 3 and in chapter 4). As in the prediction problem TD methods for policy improvement rely on dynamic programming algorithms to compute optimal policies. Given a perfect knowledge of the MDP and some candidate policy π a new policy π' can be obtained by acting greedily with respect to V^{π}:

$$\pi'(s) = \text{argmax}_{a \in \mathcal{A}} \Big[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}_{ss'}^a \tilde{V}^{\pi}(s') \Big] \tag{1.16}$$

It is easy to see that the new policy is equal or better than the old one. If the number of states and actions are finite and policy evaluation is systematically followed by policy improvement the cycle can be proven to always converges to the optimal policy π^*. This alternation of policy evaluation and improvement is known as *policy iteration* and all model-free methods rest on this procedure to find the optimal policy. In particular actor-critic methods are TD methods that have two separate structures to represent the state-value function and the policy. The policy structure is known as the actor because it selects actions. The critic element estimates the value function and provides an evaluative feedback ("criticizes") of the conduct of the actor.

These methods were first introduced in (Barto et al., 1983) where it was shown how a system consisting of two neuron-like adaptive elements can effectively solve a difficult credit assignment problem. The key idea behind these methods is to use the reward prediction error signal to simultaneously find the best policy and to estimate the state-value function for the current policy. Indeed, while interacting with the environment, the agent can evaluate how good is an action even in absence of immediate reinforcement by simply considering whether the next state is more or less attractive of the previous one in term of future reward. In this sense the TD error signal, that in absence of immediate reward is $\delta_t = \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$, represents a surrogate of the reinforcement. Therefore modifying the policy according to the TD error signal naturally leads to strengthen the probability of those actions that lead the system to states more valuable (and to decrease the probability of less appropriate actions).

Below two implementation of the actor-critic architecture will be presented, mainly following Dayan and Abbott, 2001. In both cases the state value function is updated

according to Equation 1.5, but they differ in the way the teaching signal δ_t is used to improve the policy.

Let assume that the policy is parametrized using a set of modifiable weights and that each weight $\nu(s,a)$ indicates the tendency to select (preference for) action a when the system is in the state s. These weights are then used to implement the policy via a softmax distribution:

$$\pi(s,a) = \frac{\exp(\beta\nu(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\beta\nu(s,a'))} \tag{1.17}$$

where β is a parameter that governs the exploration/exploitation trade-off. For large β the policy becomes almost a deterministic function of the preference parameters $\nu(s,a)$, if β is small the policy tends to become random.

A first way to improve the policy, known as the *indirect actor*, is to base the action choice on *state-action value function* $\tilde{Q}^\pi(s,a)$. This function is defined as the expected return when the action a is selected in the state s and the policy π is followed afterwards. The preference weight $\nu(s,a)$ approximates the value of $\tilde{Q}^\pi(s,a)$ by using the following update at each time step t[10]:

$$\nu_{t+1}(s,a) = \begin{cases} \nu_t(s,a) + \eta\delta_t & \text{if} \quad s_t = s\,,\ a_t = a \\ \nu_t(s,a) & \text{if} \quad s_t = s\,,\ a_t \neq a \end{cases} \tag{1.18}$$

where η represents the learning rate for the actor element.

An alternative way to learn the policy, known as the *direct actor*, consists in updating the preference directly to maximize the expected long-run return. This can be done by stochastically following the gradient $\partial\tilde{V}^\pi(s)/\partial\nu(s,a)$. Such a gradient can be written as:

$$\begin{aligned} \frac{\partial\tilde{V}^\pi(s)}{\nu(s,a)} \quad \propto \quad & \pi(s,a)\big[1-\pi(s,a)\big]\Big(\mathcal{R}_s^a + \gamma\sum_{s'\in\mathcal{S}} \mathcal{T}_{ss'}^a \tilde{V}^\pi(s') - C(s)\Big) - \\ & - \sum_{a'\neq a} \pi(s,a')\pi(s,a)\Big(\mathcal{R}_s^{a'} + \gamma\sum_{s'\in\mathcal{S}} \mathcal{T}_{ss'}^{a'} \tilde{V}^\pi(s') - C(s)\Big) \end{aligned} \tag{1.19}$$

where a state dependent constant $C(s)$ have been included in both terms on the right side because it cancels out. After starting from the state s_t the experienced immediate reward r_{t+1} and the value of the next state $V^\pi(s_{t+1})$ can be used as a guess for the corresponding expected value and setting the constant $C(s) = V^\pi(s)$ the preferences can be updated as:

$$\nu_{t+1}(s,a) = \begin{cases} \nu_t(s,a) + \eta\delta_t\big[1-\pi(s,a)\big] & \text{if} \quad s_t = s\,,\ a_t = a \\ \nu_t(s,a) - \eta\delta_t\pi(s,a) & \text{if} \quad s_t = s\,,\ a_t \neq a \end{cases} \tag{1.20}$$

[10]This is because the state-value function can be written as $\tilde{V}^\pi(s) = \sum_{a\in\mathcal{A}} \pi(s,a)\tilde{Q}^\pi(s,a)$. The agent experiences a sample of $\tilde{Q}^\pi(s,a)$ each time it visit the state s and selects the action a. Therefore, if the policy changes slowly, the preference weight $\nu(s,a)$ tracks the value of $Q^\pi(s,a)$.

Both the Equation 1.18 and Equation 1.20 combined with TD methods for policy evaluation represent a stochastic version of the policy iteration algorithm. The Actor/Critic algorithm has not been proved to converge because policy evaluation and policy improvement are carried out simultaneously. Nevertheless, it has been the framework most widely used as a model of neural reinforcement learning and action selection in the basal ganglia (Barto et al., 1983; Barto, 1995; Houk et al., 1995; Suri and Schultz, 1999; Suri, 2002). This is mainly due to the resemblance of the Actor/Critic architecture with the anatomical division of the striatum in dorsal and ventral part (supposed to implement respectively the actor and the critic element (O'Doherty et al., 2004; see Joel et al., 2002 for a review).

A couple of model-free methods that use the state-action value function to improve the policy will be quickly sketched to conclude this section. First, it is possible to write a Bellman equation for the function $Q^\pi(s,a)$:

$$\tilde{Q}^\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(s',a') \tilde{Q}^\pi(s',a') \tag{1.21}$$

With the same reasoning employed to develop a TD algorithm for the value function it is possible to use samples of observed states and action to define a TD update for the state-action value function:

$$Q^\pi(s_t,a_t) = Q^\pi(s_t,a_t) + \alpha[r_{t+1} + \gamma Q^\pi(s_{t+1},a_{t+1}) - Q^\pi(s_t,a_t)] \tag{1.22}$$

where the function $Q^\pi(s,a)$ indicates the current estimate $\tilde{Q}^\pi(s,a)$. The algorithm described in Equation 1.22 is known as *SARSA* because it uses the sequence of state-action-reward-state-action to carry out each backup. Alternatively the agent can try to learn directly the optimal state value function $\tilde{Q}^*(s,a)$ using an off-policy TD algorithm. This reinforcement learning technique is known as *Q-learning* and uses the estimate of the optimal state-action value function to perform the backup:

$$Q^\pi(s_t,a_t) = Q^\pi(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_{a'} Q^\pi(s_{t+1},a') - Q^\pi(s_t,a_t)] \tag{1.23}$$

The above equation converges to $\tilde{Q}^*(s,a)$ because it directly implements the Bellman optimality equation for the state-action value function (Sutton and Barto, 1998).

1.1.4 TD learning with function approximation

So far the discussion focused on TD methods for table-lookup representation of the the state-value function, i.e. when states are described via localist representations (one unit per state). Here it is described how these methods can be extended when arbitrary

function approximation schemes are used to represent the state-value (also referred as distributed representation of the state).

When the state space of the MDPs increases the problem of prediction becomes unmanageable: in addition to the memory problem related with the large number of state values to store, solving the prediction problem for each state individually becomes too slow. To solve the problem for large MDPs it is suitable to replace the table representation for the value function with a parametrized functional form:

$$V(s, \mathbf{w}) \approx \tilde{V}^{\pi}(s) \tag{1.24}$$

where $\mathbf{w} \in \mathbb{R}^n$ indicates a vector of parameters that are adjusted during the learning process using TD methods. The number of parameters n is typically much less than the number of states and therefore function approximation reduces the computational cost of the problem and more importantly allows to generalized from experienced state to ones that have never been seen. Generally the performance of function approximation method is measured using a cost function defined as expected squared difference between the approximate value function $V(s, \mathbf{w})$ and the true value function:

$$J(\mathbf{w}) = \frac{1}{2} E^{\pi}\Big[\big(\tilde{V}^{\pi}(s) - V(s, \mathbf{w})\big)^2\Big] \tag{1.25}$$

The cost function $J(\mathbf{w})$ can be minimized using stochastic gradient descent methods, i.e we can apply use gradient descent methods to the experienced samples to find a local minimum. According to this approach the parameter vector $\mathbf{w}$ is updated after each experienced state s_t in the direction that would reduce the squared-error on that specific sample:

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} J(\mathbf{w}_t) \\ &= \mathbf{w}_t + \alpha\big[\tilde{V}^{\pi}(s_t) - V(s_t, \mathbf{w})_t\big] \nabla_{\mathbf{w}} V(s_t, \mathbf{w}_t) \end{aligned} \tag{1.26}$$

where α is a small parameter defining the learning rate. In the function approximation scheme states are represented by feature vectors which component provide a compressed characterization of the states. Hereafter only linear function approximation will be discussed because stochastic gradient descent methods are proven to converge to some estimation of $\tilde{V}^{\pi}(s_t)$ (whose error can be quantified) in this case (Bertsekas and Tsitsiklis, 1996). Let then write the state value function at time t as $V(s_t) = \mathbf{x}_t \cdot \mathbf{w}_t$, where $\mathbf{x}_t = \{x_t^1, x_t^2, \cdots x_t^n\}$ indicates a n-component feature vector. Following the same reasoning of the previous sections the value $\tilde{V}^{\pi}(s_t)$ can be estimated through the observed

samples. Using as estimate the 1-step return and according to (1.26) we can define the TD(0) backup for linear value function approximation as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t\mathbf{x}_t \tag{1.27}$$

Stochastic gradient descent methods can be similarly applied to implement the TD(λ) backup after introducing an appropriate eligibility trace for the feature vector $\mathbf{x}$. Following (Sutton and Barto, 1998) the temporal evolution of $\mathbf{e}_t$ for accumulating traces can be defined as:

$$\mathbf{e}_{t+1} = \lambda\gamma\mathbf{e}_t + \nabla_{\mathbf{w}}V(s_t, \mathbf{w}_t) = \lambda\gamma\mathbf{e}_t + \mathbf{x}_t \tag{1.28}$$

where the last equivalence holds because the attention has been restricted to linear function approximation. The temporal evolution of replacing traces and for linear approximation can be written as:

$$e^m_{t+1} = \begin{cases} x^m_t & \text{if} \quad x^m_t \neq 0 \\ \gamma\lambda e^m_t & \text{otherwise} \end{cases} \tag{1.29}$$

where e^m_t indicates the m-component of the vector $\mathbf{e}_t$. In both accumulating and replacing traces at each time step t the TD(λ) update for the weight vector $\mathbf{w}$ is defined as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t\mathbf{e}_t \tag{1.30}$$

From the above equations it is easy to see that TD methods with the lookup-table are just a special case of the TD algorithm with linear function approximation. When the feature vector is defined using the *presence representation*, i.e. it is defined as $x^m_t = 1$ if $s_t = s_m$ and 0 otherwise the (1.30) reduces to (1.13). Under some technical assumptions on the decrease of the learning rate stochastic gradient descent TD(λ) methods are proven to converge close to a global optimum (i.e to a global minimum of the error function) when the function approximation is linear (this is because the cost function is a quadratic function of the parameter vector $\mathbf{w}$; see Tsitsiklis and Van Roy, 1997 for details).

1.1.5 Partial observability

The reinforcement learning formalism described in the previous sections presents several difficulties when applied to real world situations. One crucial lack of the approach discussed so far is the assumption of fully observability: in MDPs, indeed, the agent does have full access to all the information necessary to precisely describe the situations it comes across, i.e. the agent directly observes the environment state. The abstract description of the agent environment interaction assumed in MDPs is clearly faraway from

the context where realistic decision making processes happen. Generally the agent does not observe the environment state directly but it needs to make decisions relying on noisy sensory informations that describe the environment only partially. Decision making under uncertainty can be appropriately described as a partially-observable MDP (POMDP, see Kaelbling et al., 1998, for background) that is a process whose dynamics abides by the Markov property, but whose states cannot be observed directly. Instead of directly experiencing the state of the world, at each time step, the agent receives an observation o_t drawn from a state dependent probability distribution $\mathcal{O}_s = P(o_t|s_t = s)$. Different reinforcement learning approaches can be used to cope with the problem of partial observability. Here special importance is given to a Bayesian probabilistic approach because it has been suggested that when the knowledge of the world is incomplete the brain implements Bayesian inference during perception (Rao et al., 2002; Knill and Richards, 1996). In the Bayesian scheme states are replaced by beliefs, i.e. posterior probability about the state of the environment, that can be then combined with reinforcement learning methods to select optimal actions (Dayan and Daw, 2008; Bogacz and Larsen, 2011). A belief state is a probability distribution over a set of hidden states $\mathcal{S}$ that depends on the history of the system $H_t = o_t, a_{t-1}, o_{t-1}, \ldots, a_0, o_0$ (i.e all the past observations and actions). The ith component of the belief state is defined as $b_t(i) = P(s_t = i|H_t)$ and can be computed recursively over time from the previous belief state using Bayes rule:

$$\begin{aligned} b_t(i) &= k \cdot P(o_t, |s_t = i, a_{t-1}, o_{t-1}, \ldots, a_0, o_0) P(s_t = i|a_{t-1}, o_{t-1}, \ldots, a_0, o_0) \\ &= k \cdot P(o_t, |s_t = i) \sum_j P(s_t = i|a_{t-1}, o_{t-1}, s_{t-1} = j) P(s_{t-1} = j|a_{t-1}, o_{t-1}) \\ &= k \cdot P(o_t, |s_t = i) \sum_j \mathcal{T}^a_{ij}\, b_{t-1}(j) \end{aligned} \tag{1.31}$$

where k is a normalization constant and the Markov property has been used. The fact that the belief can be updated recursively using the last observation and action implies that the belief state $\mathcal{B}$ is a sufficient statistic, i.e in the belief state the process is a MDP. Offline exact solutions for POMDPs in the finite horizon case have been discussed in (Kaelbling et al., 1998). In general once converted in a MPD offline solution for the POMDP can be find using standard dynamic programming algorithm (see for example Sutton and Barto, 1998) and this approach have been successfully applied to the study of reward optimization decision making under uncertainty in the brain (Huang and Rao, 2013). Considering that in the belief space $\mathcal{B}$ the process is an MDP one can also use function approximation techniques to solve the POMDP with standard online methods for MDPs. Relying on this consideration Rao, 2010 suggested a possible implementation of

neural decision making under uncertainty in the brain using an actor-critic architecture and TD learning. The ideas and the algorithm suggested in that work will be briefly sketched in subsection 1.3.3.

1.2 Behavioural experiment and neurophysiology of the dopamine system

This section describes the neurophysiology of reward structures in the brain, mainly focusing on the activity of the dopamine-producing neurons, a small group of neurons located in two areas of the midbrain, the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc). Dopamine is one of the four mayor neuromadulators of the brain and the dopaminergic system is involved in a variety of important brain functions such as reward-dependent learning, motivation, motor control. The loss of a large number of dopamine neurons leads to Parkinson's disease and in general the dysfunction of the dopaminergic system is associated with a number of disorders including addiction, schizophrenia, and ADHD.

The discussion will focus on the role of dopamine in reward-related learning because of its relationship with computational theory of reinforcement learning described so far and in particular because of the close parallel between TD learning and the activity of dopamine producing neurons. This parallel is expressed by the reward prediction error hypothesis of dopamine, stating that the phasic activity of dopaminergic neurons codes a discrepancy between the predicted and currently experienced rewards,a pattern of activity that strictly resembles the reward prediction error signal of the TD learning algorithm (see Equation 1.5).

Further evidence emphasizing the role of dopamine in reward-related learning rests on the strong interaction between dopamine neurons in the midbrain and the striatum, a critical component of the motor and reward system. The striatum is the main input structure for a group of subcortical nuclei situated at the base of the forebrain known as the basal ganglia. It receives organized, convergent inputs from widespread areas of the cortex, which are projected to other basal ganglia structures such as the globus pallidus, and eventually projected back to cortex through the thalamus. Dopaminergic neurons crucially affect the activity of the basal ganglia circuits because of their interaction with striatal neurons. They receive inputs that are related to primary reinforcement of appetitive nature coming from different brain areas (such as the pedunculopontine

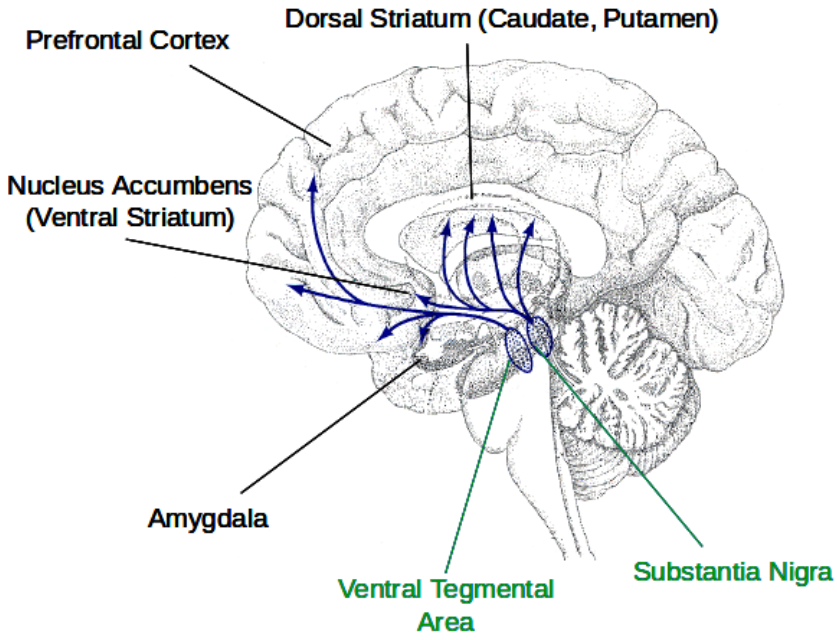


Figure 1.1: Schematic view of the dopamine pathways.

nucleus, the habenula, the amygdala, ecc.) and inputs from striatal neurons that are related to cortical activity. Crucially they project back their reward prediction error signal to the same striatal zone that send them inputs. Specifically neurons in the SNc mainly project to the dorsal parts of the striatum, and those in the VTA to the ventral striatum (and to the prefrontal cortex; see Figure 1.1). Additionally, stiatal areas have been demonstrated to be subject to dopamine-dependent long-term potentiation (LTP) and long-term depression (LTD) (Wickens et al., 1996; Calabresi et al., 2000; Reynolds and Wickens, 2002).

All these results suggest very clearly that the phasic activity of the midbrain dopamine neurons provides a global teaching signal to the striatum that guides reward related learning through a mechanism of dopamine-dependent plasticity in corticostriatal synapses (Schultz, 1998). These synaptic modifications and the reward prediction error signal of dopamine neurons, in turn, provide the mechanistic underpinning for a specific class of reinforcement learning mechanisms in the brain, and suggest that humans and many other animals may be implementing algorithms that correspond in a close way to the algorithms of the formal theory.

In what follows data and theory about relevant aspects and function of the dopamine system will be reviewed. The attention will focus on the main steps that have led to the formulation of the reward prediction error hypothesis and that support this theory. This

approach is followed for two reason: on the one hand, although alternative hypothesis about what type of signal is conveyed by dopaminergic activity in the midbrain exist (see for example Berridge, 2007; Friston et al., 2012), and the type of information coded by dopaminergic transmission is very likely to go beyond reward-prediction error, it is widely accepted that "to date no alternative has mustered as convincing and multidirectional experimental support as the prediction-error theory of dopamine" (Niv and Montague, 2008; but see also Glimcher, 2011). On the other hand, in the central part of this thesis (chapter 3 and chapter 4) I will assume the validity of this hypothesis and study how Bayesian inference and RL algorithms can be integrated to improve our understanding of the dopamine neurons activity when animals need to rely their decision on uncertain stimuli.

Before proceeding with the presentation of some relevant electrophysiological findings about the dopamine activity the learning paradigm of conditioning will be briefly introduced. Conditioning is relevant in what follows for two reasons: the majority of existing data about dopamine neurons have been recorded during behavioural conditioning tasks. And the TD algorithm was initially developed to explain behaviour and only during the last two decades it became a leading model of the brain.

1.2.1 Conditioning

Conditioning is the study of how animals learn to predict appetitive outcomes and direct their behaviour accordingly. It is common to make a distinction between two forms of conditioning: Pavlovian (or classical) conditioning, and instrumental (or operant) conditioning. In the former outcomes occur regardless of the animal's actions and learning consists in acquir the ability to predict the causal relationships between events in the world. In the latter the animal directly affects the occurrence of significant outcomes and learn to strengthen those actions that lead to positive consequences.

In the typical Pavlovian paradigm an initially neutral stimulus (known as conditioned stimulus, CS) is associated with a biologically important stimulus (unconditioned stimulus, US) by repeatedly present the two stimuli with some predefined temporal relationship. The temporal interval between the CS (usually a visual cue) and the US (normally a drop o juice that serves as reward) is denominated inter-stimulus interval (ISI); whilst the temporal interval separating two consecutive trials is known as inter-trial interval (ITI). The CS need to be presented before than the US in order to establish a causal relationship between the two stimuli (i.e the ISI must be positive). In the *delay conditioning* procedure the CS is presented for some period of time and the US is delivered when the CS

Phenomenon	Pre-Train	Train	Result
Acquisition		A→US	A $\Longrightarrow$ CR
Extinction	A→US	A→ $\cdot$	A $\Longrightarrow$ $\cdot$
Partial		A→US ($P_{A,US}$) A→ $\cdot$	A $\Longrightarrow$ α_PCR
Blocking	A→US	A,B→US	A $\Longrightarrow$ CR B $\Longrightarrow$ $\cdot$
Overshadow	A→US	A,B→A	A $\Longrightarrow$ α_ACR B $\Longrightarrow$ α_BCR
Secondary	A→US	B→A	B $\Longrightarrow$ CR

Table 1.1: Some major effects in classical conditioning. Train and Pre-Train typically consists in several repetitions of each trial type. A,B indicates two different CS's. In the "Train" column the symbol "$\cdot$" indicates omission. In the "Result"column the "$\cdot$" denotes a CR that is missing (or significantly diminished), the factors of α denote a partial or weakened expectation. The α_P in the case of partial reinforcement indicates that the partial expectation depends on the probability $P_{A,US}$ of reward in the training stage.

is still present. When instead the CS and the US are separated by a temporal delay the procedure is known as *trace conditioning*. In both cases, after pairing is repeated, the (initially neutral) CS alone starts to elicit a behavioural response, known as conditioned response (CR), that is similar to the one naturally caused by the US. The association is reached faster in the delay procedure (because of the temporal contiguity) and the degree of the CR typically reflects the degree of association between the two stimuli (i.e the degree to which the US is expected due to the presentation of the CS). In addition to the appearance of the CR a variety of behavioural effects have been observed in classical conditioning. If the US is repeatedly omitted after the presentation of the CS the association between the two stimuli and consequently the CR disappear. This phenomena is known as *extinction*. Another effect known as *blocking* consists in the observation that if an US is already predicted by a CS then adding a second CS that also predict the US produces only reduced learning. A summary of some major effects observed in classical conditioning can be found in Table 1.1.

Instrumental conditioning includes three types of experiments: *free operant task*, and *discrete choice tasks*. In free operant tasks an animal is placed in a computer-controlled box (known as operant conditioning chamber or "Skinner Box") and learns to respond to specific stimuli (such as a light or a sound signal) with given actions (like pressing a lever). Correct performance is associated with rewards delivery, and, in some cases, the animal receives a punishment for incorrect or missing responses. Free operant tasks can follow two main different programmed schedules, the ratio schedule and the interval

schedule. In the ratio schedule, one out of n leverpresses is rewarded. The number of leverpresses required for the reward delivery can be constant or variable across trails. In the interval schedule, for instance, a reward occurs following the first leverpress after an interval of time has elapsed from the previous reward. The threshold interval can be fixed or variable. A classical result in the fixed interval schedule is that the leverpressing sharply increases when the end of the fixed interval is approaching. This schedule has been recently used to study how striatal neurons are involved in timing (Mello et al., 2015) and I will discuss significant findings in the next chapter. Discrete choice tasks consist in a two-armed bandit task: the animal has to decide between two different stimuli associated with different reward magnitude, delay and probability. This type of experiments have been extensively used to investigate time discounting. The way how animals and humans are discounting future rewards cannot be measured directly. However key insights on this issue can be obtained by observing how their preferences change when reward magnitude and delay vary. Given that an animal is presented with two different amounts of reward, r_1 and r_2, delivered after different delays, d_1 and d_2, and that it systematically prefers one of them, if the delay of both rewards is increased of the same temporal interval Δ the preference should not change under the assumption of exponential discounting. Actually in many species of animals, from pigeons to primate (and to humans), the preference does change depending on the duration of the interval Δ. This behavioural effect is known as *preference reversal* and can be justified under the assumption of hyperbolic discounting. Although in the rest of thesis I will formulate the reinforcement learning problem in the term of exponential discounting it is interesting to notice that some evidence for hyperbolic discounting have been observed even in the response of dopamine neurons (Kobayashi and Schultz, 2008). However the result obtained in that work was not completely clear and the distinction between hyperbolic and exponential discount models at single neurons level was not striking.

1.2.2 Dopamine activity reward prediction errors

The study of the dopamine activity initially focused on the role of dopamine in motor control. This was due to the fact that one relevant symptom of Parkinson's disease, known to be caused by the degeneration of dopamine neurons, is the loss of motor control. Surprisingly early recording conducted by Romo and Schultz, 1990 showed that movement alone did not imply any response, and that the response of dopamine neurons was unrelated to movement parameters, but rather it was signaling an expectation of reward. More specifically Romo and Schultz, 1990 trained two monkeys to perform two different

tasks: in the first experiment the monkeys performed self-initiated arm movements from a resting key into a covered box containing food. In a second task, the arm movement was triggered by the rapid opening of the door of the food box. They found that neither in the first nor in the second task the dopamine neurons activity was correlated with the initiation of the movement. In the first task they observed a significant response after the monkey's hand touched the food inside the box or the bare wire normally holding the food. Importantly they found that the touching of the same bare wire outside of the behavioural task did not elicit any response. In the second task they observed that, after some period of training, neurons started to respond mainly to the door opening and not to the food. These findings constituted the first steps toward the reward prediction error hypothesis and inspired series of later experiments.

Subsequent works from the same group focused on analysing the activation of dopamine neurons during different steps of learning of simple instrumental conditioning reaction time tasks (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz and Schultz, 1994). These studies highlighted the importance of unpredictability for reward responses in dopamine neurons showing that when learning was established the response of neurons transferred from the reward to the reward-predicting, movement-triggering stimulus. In addition, in trials in which the trained monkey received no reward (because the key pressed was accidentally wrong) many of the dopamine neurons showed a sharp decrease in their firing rates below baseline shortly after the reward's usual time of delivery. These results suggested that dopaminergic activity is sensitive to both the occurrence and time of the reward. The idiosyncratic pattern of dopaminergic activity observed during these experiments is summarized in Figure 1.2.

Early electrophysiology studies guided the formulation of the reward prediction error hypothesis and of the first TD-based model of the dopamine system (Montague et al., 1996; Schultz et al., 1997). The model proposed by Montague et al., 1996 elucidated how and why the normative approach of reinforcement learning and in particular the TD algorithm accounted (qualitatively) for the output of dopamine neurons during learning. It provided strong evidence to the hypothesis that the output of these neurons was consistent with the scalar prediction error signal required by the normative reinforcement learning models. The details of this model implementation together with the way how the mismatches between the TD error resulting from the original model and dopamine neuron activity have been be addressed will be discussed later in this chapter.

Many other studies have investigated weather the dopamine responses had the characteristic of the teaching signal of the formal theory and weather the signal conveyed by

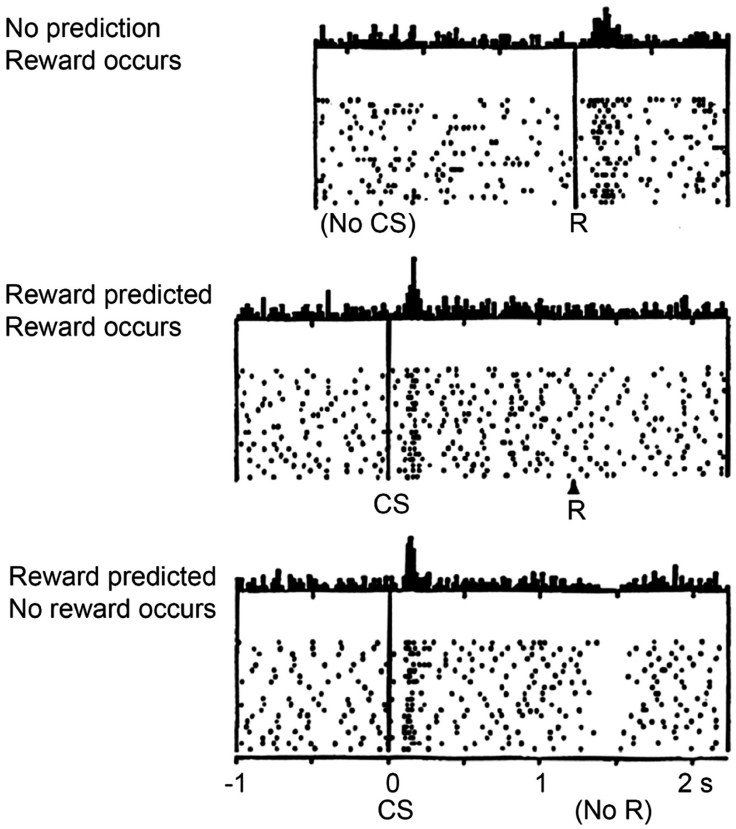


Figure 1.2: Standard phasic activation of a typical dopamine neuron resembling an error in reward prediction, from (Schultz et al., 1997).

this group of neurons reflected learning at behavioural level. Hollerman and Schultz, 1998 found that in a discrimination task the response to the reward progressively decreased in a way consistent with the course of learning at behavioural level. Waelti et al., 2001 using a blocking procedure confirmed that learning both at neural and at behavioural level depended crucially on prediction errors and not on stimuli-reward association alone.

During the last decade the validity of the reward prediction hypothesis has been confirmed by a wide number of quantitative tests. Using using a general regression model Bayer and colleagues showed that the dopaminergic response to the reward in the current trial could be predicted using an exponentially recency weighted average of previously experienced rewards (this is implicit in the TD update as discussed in subsection 1.1.2). Additionally this study showed that, at least in the positive domain, it was possible to extract (positive) reward prediction errors from dopamine firing rates. Later it was shown that negative prediction errors were encoded by the length of the pause below the baseline firing rate, rather than the magnitude of the pause (Bayer et al., 2007).

Further works manipulated the magnitude and/or probability of the reward occurrence

and showed that the phasic dopaminergic response to the CS was in agreement with TD models predictions and resulted proportional to the magnitude and/or probability of the predicted reward (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005). (Fiorillo et al., 2003) analysed the response of neurons in a delay conditioning experiment in which reward was delivered only probabilistically. They encountered that the population response averaged in trials showed a clear ramping profile during the interval between the CS and the reward, and that this ramping activity seemed to reflect uncertainty in the reward delivery, rather than a prediction error. Niv et al., 2005 reconciled this experimental finding and TD models showing that, if negative and positive prediction errors were encoded differentially in the dopamine neurons (as shown in Bayer and Glimcher, 2005), the ramp was in accord with a constantly back-propagating error signal as predicted by the existing formal theory.

The majority of the studies mentioned above used rather simple Pavlovian or instrumental tasks. A first analysis of the dopamine responses in a more complex scenario was reported in (Morris et al., 2006). Monkeys were trained to perform an instructed-choice task and a two-armed bandit choice task. They found that, in choice trials, the response of neurons immediately after the presentation of the two cues reflected the value of the subsequent choice, a pattern of activity in line with SARSA learning. Another study performed in rats (Roesch et al., 2007) showed that the activity of dopamine neurons complied with the predictions of Q-learning. These contrasting results can be due to many differences between these two studies, including the animal species (a tendency for optimal and suboptimal behaviour have been reported respectively in rats and monkeys). Finally, it is worth to note that both these results are at odds with straightforward predictions of an Actor/Critic mechanism (although this will be main action selection mechanism used in this thesis).

Recently, there has been an increasing interest in investigating the response of dopamine neurons in more realistic decision making tasks (Nomoto et al., 2010; de Lafuente and Romo, 2011). The existing studies seem to indicate that the DA signal has a much richer structure than in simple choice paradigms. For example, Nomoto et al., 2010, found that the response to visual dynamic random dot stimuli is more complex than the response to the stimuli commonly used in previous studies. The DA activity seemed to follow a more elaborate temporal profile, first responding abruptly to the onset of the stimulus (presumably due to its detection) and then producing a more extended response (supposedly due to the decision-making process, Nomoto et al., 2010 ; see also Schultz, 2015). In another recent study, de Lafuente and Romo, 2011 recorded DA neurons while a monkey was

engaged in the detection of weak vibrotactile stimuli. In this task, when the animal was instructed to communicate its choice by pushing one of two push buttons, these neurons responded with a burst that coded the animal's uncertainty on its own judgement. The activity of dopamine neurons in two specific decision making tasks will be investigated in chapter 3 and in chapter 4. In particular, in chapter 3 I will perform a detailed model based (re-)analysis of that data reported in (de Lafuente and Romo, 2011).

Another issue that has been only glanced until now concerns the relationship between dopamine responses and the processing of time (see Daw, 2003 for an extensive review on behavioural data and early electrophysiological recordings about this topic).

Schultz et al., 1993 were the firsts to notice that, the responding of dopamine neurons appeared to be high sensitive to the exact timing of relevant task events and strictly related with the temporal predictability of their appearance. They trained a monkey to perform three different versions of an instrumental conditioning task. In the experiment one of two light instructions and a trigger were turned on at different delays from each other and the animal had to wait the trigger presentation before responding accordingly to the instruction previously presented in order to receive a reward. After sufficient training the monkey was able to correctly perform and, as previously reported, after learning neurons ceased to respond to the reward delivery and started to activate after the instruction presentation. In addition to a pause at the time of the expected reward in occasionally wrong trials, the author noticed another additional remarkable feature in the dopamine response. They found that neurons did not respond to the trigger when the instruction-trigger delay was fixed but they did respond when the delay was variable (i.e respectively in the second stage and third stage of the task; see Figure 1.3, left). In another study (Hollerman and Schultz, 1998) a monkey was trained to expect a reward one second after a cue, and the authors analysed the responding of neurons in occasional probe trials in which rewards were moved up or delayed (of half second in both cases). If a reward was delayed (similarly to what happened when it is omitted altogether), dopamine neurons showed a pause in their background firing at the time when the reward should have occurred and a phasic burst at its (delayed) delivery. An early reward caused a dopaminergic burst, but there was no corresponding pause at the time the reward was normally delivered (see Figure 1.3, right). The observations described above suggested that the monkeys were internally keeping track of the timing of the relevant task events and that these temporal expectations were reflected in the activity of dopamine neurons.

More recently Fiorillo et al., 2008 analysed more in detail how the activation of dopamine neurons related with temporal expectations. They found that both at be-

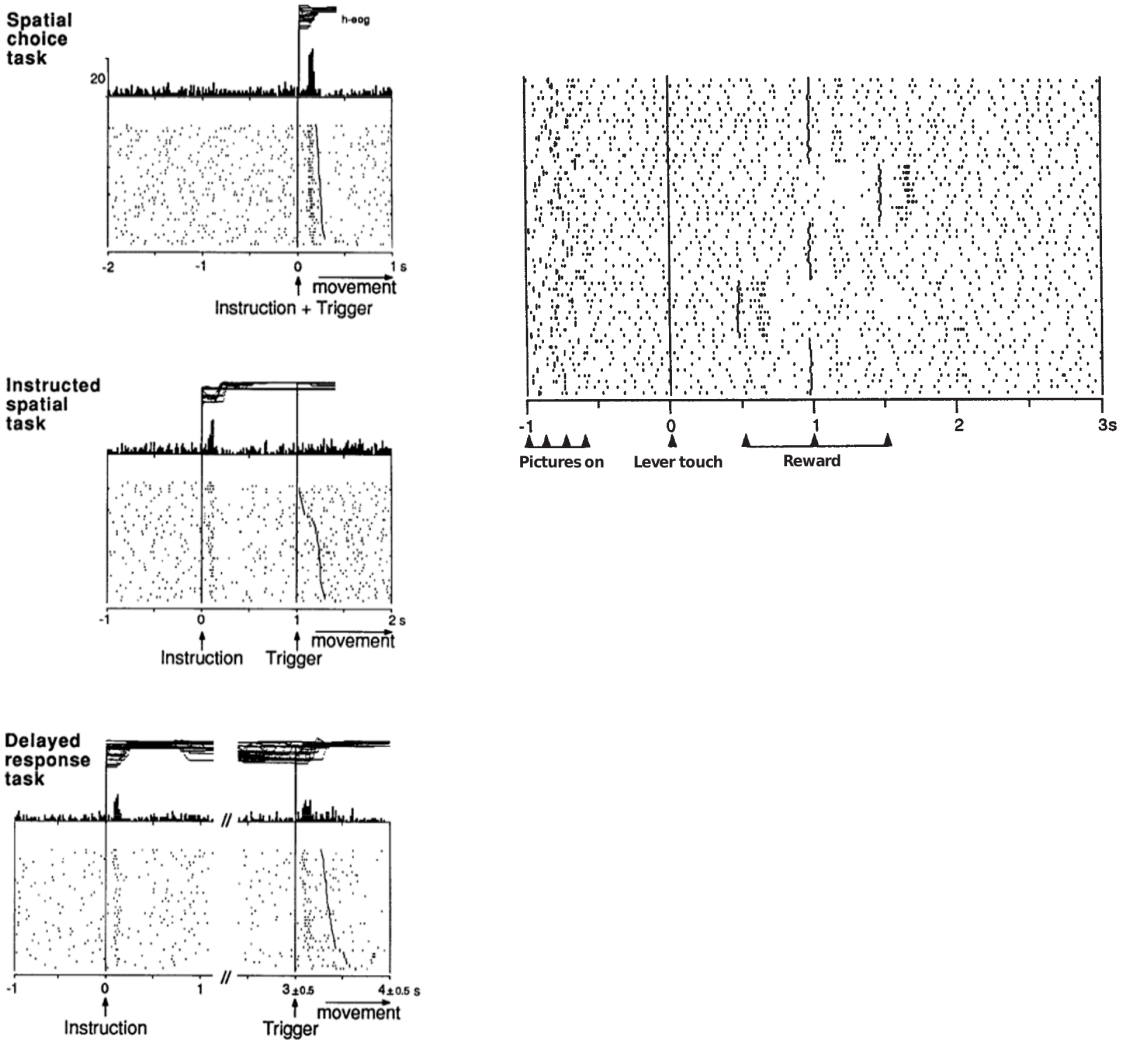


Figure 1.3: Response of dopamine neurons to timing varying events. Left: In the first version of the task ("Spatial choice task") the trigger and the instruction were presented together, and, when learning is established, neurons respond to their occurrence. In the "Instructed spatial task" the instruction-trigger delay was fixed to 1 s and the trigger did not generate any response because it was fully predicted by the instruction. The response to the trigger persisted, although learning was fully established, when the instruction-trigger delay varied between 2.5 s and 3.5 s ("Delay response task"). Data from (Schultz et al., 1993)). Right: Effects of reward timing on dopamine activation. During familiar trials, following a correct response, the reward was delivered after 1.0 s. In occasionally probe trials the reward was delivered after 1.5 s (delayed) or after 0.5 s (early). Data from (Hollerman and Schultz, 1998).

haviour and at neural level the precision of the expectation was not high and that the

precision declined with the passage of time in way consistent with the Weber's law. Similar results were found in (Kobayashi and Schultz, 2008). Additionally it has been shown that the presence of time-varying reward related events produced a tonic slow decrease of the dopamine neurons firing rate. The deviation from the baseline started from the first time the event could be presented and increased in magnitude as time elapsed resembling a form of negative prediction error (Bromberg-Martin et al., 2010; but see also Fiorillo et al., 2008; Nomoto et al., 2010; Pasquereau and Turner, 2015; Starkweather et al., 2017 for similar results). More interestingly, it has been recently proved that dopamine neurons are indeed implicated in the perception of the passage of time (Soares et al., 2016). Another quite new study analysed the response of dopamine neurons in a classical conditioning task in which the reward, in addition to be presented after a variable interval from the CS, was delivered only in the 90 % of the trials. The main result was that dopamine neurons reflected prediction errors provided that the TD machinery received a hidden variable reflecting Bayesian inference as input.

Both Bayesian inference and the representation of elapsed time in TD models of the dopamine activity will be central to this thesis. They represent fundamental elements to take into account when TD algorithms are used to model task in which decisions rely on temporal and sensory uncertain stimuli. In particular chapter 2 will provide a more detailed description of the relationship between the firing rate of dopamine neurons and temporal expectation and I will analyse in details the way a particular TD model deals with current data. The necessity to incorporate Bayesian inference in TD models of the dopamine system will be addressed in depth in chapter 3 and in chapter 4 where I will discuss a model based analysis of data from two specific decision making tasks.

1.3 TD models of the dopamine response

This section will conclude the literature review with a brief discussion of existing TD model of the dopamine activity. The discussion will mainly focus on the representation of time in TD model of the basal ganglia (a topic that will be resumed in the next chapter). The way Bayesian inference and TD learning have been integrated in a quite recent work (Rao, 2010) will be also discussed.

1.3.1 Early TD models

The most specific and now standard formulation of TD model of the dopamine system was introduced in (Montague et al., 1996) (presented here as in Schultz et al., 1997). In that model each task event activated a a set of N features $\mathbf{x}_t = \{x_1(t), x_2(t) \dots x_N(t)\}$ corresponding the animal's representation of the temporal interval elapsed from the event onset. Using the terminology of section 1.1 the state of the system corresponded to the interval of time between relevant task events(in the simplest case just between the CS and the reward). The value of each event was estimated as a linear combination of these features $V(s_t) = \mathbf{w}_t \cdot \mathbf{x}_t$, and events were supposed to contribute in an additive way to the value function. The value function was learned by adjusting the weights according to the TD(0) algorithm (see Equation 1.27). This early application of TD learning to the dopamine system assumed a tapped-delay line representation of stimulus history, also known as the complete serial compound (CSC; Sutton and Barto, 1990). Time was discretized into time steps of a typical duration of 100 ms (a timescale corresponding to the response latency of dopamine neurons; see for example Romo and Schultz, 1990; Schultz and Romo, 1990), and the CSC represented every time step following stimulus onset as a separate feature, making implicitly the assumption of a perfect clock. This assumption is clearly unrealistic and produces a series of erroneous predictions that will be discussed below and in particular chapter 2. Another noteworthy simplification was that the model considered the return being predicted as confined to single trials. The model correctly reproduced that the response of dopamine neuron activity shifted from unpredicted rewards to the predictor; that when more than one predictor was presented neuron responded to a reward-predicting stimulus only when its timing relative to a previous predictor was variable (Schultz et al., 1993). The predictions of the model qualitatively mismatched with the data when the reward was occasionally omitted, showing a too pronounced and short pause. More importantly the model also mispredicted part of the results of the Hollerman and Schultz (1998) experiment, in which a reward occasionally occurred earlier or later than expected. Predictions qualitatively corresponded to the data when the reward was delayed with respect to its usual time of occurrence; however when the reward was presented early, the model predicted a pronounced pause at the time when the reward was expected. which was not observed in the activity of dopamine neurons. To conclude, two issues that will be discussed in detail in chapter 2 deserve to be mentioned. First, the model generates an erroneous prediction of the response of dopamine neurons when the interval between relevant task events varies accordingly to some predefined schedule (for example the ones reported in Fiorillo et al., 2008; Bromberg-Martin et al., 2010). Second,

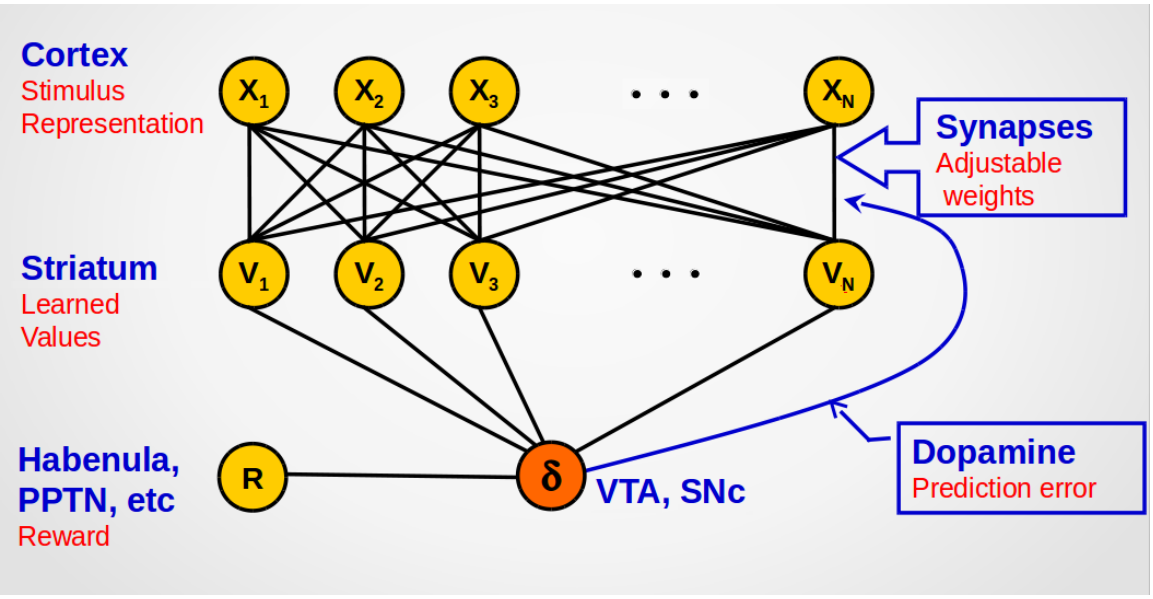


Figure 1.4: Schematic view of possible implementation of TD learning in the costricostriatal circuits.

due the unrealistic representation of time the model is also unable to deal with all those aspect of dopamine activation related to timing noise.

Many refinement have been added to the original theoretical formulation and will be discussed below. However no one of these changes questioned the core idea behind the reward prediction error hypothesis of dopamine. Besides establishing the parallel between dopamine responses and TD errors Schultz and colleagues were pioneers in linking the computational theory with the neurobiology of the brain. They envisioned that the targets of the VTA dopamine neurons (e.g. amygdala and ventral striatum) were involved in learning underlying value prediction, and that dopamine projections from SNC to dorsal striatum were responsible for behavioural control. The neural basis of reinforcement learning have been widely investigated during the last two decades.

According to RL models of the basal ganglia, states are represented in the cortex and act as cortical inputs to the striatum. Specifically, in the model proposed in (Schultz et al., 1993) these cortical inputs correspond to the set of features used to represent the stimulus and to linearly estimate the value function. The striatum itself (in particular the ventral striatum) is assumed to encode the estimated value. The adjustable weights of the TD algorithm represents corticostriatal synapses. The role of dopamine is to modulate corticostriatal plasticity similarly to how the temporal difference signal modifies the set of weights in the algorithm. A schematic view of the correspondence between reward-based prediction learning in the brain and the TD algorithm of RL is depicted in Figure 1.4.

A subsequent model was introduces in Suri and Schultz, 1998, 1999. The model was conceived to show that an actor-critic architecture could learn a spatial delayed response task as the one in (Schultz et al., 1993). Such a model resembled that of (Montague et al., 1996) in many aspects, but used a different temporal representation and a variation on the usual TD algorithm to update the weights. Specifically they used a large number of temporally extended representational elements that activates simultaneously at the onset of the stimulus. The value of these basis functions exponentially ramped up with the elapsed time and each element abruptly decreased its value to zero at a given time step after the stimulus onset. The only weight eligible for learning was the one corresponding to the representational element that had just shut off. Most importantly the authors introduced a reset mechanism allowing a high value reward-predicting stimulus or the reward itself to erase the representation of previous stimuli. With this additional mechanism the model could reproduce the activity of dopamine neurons after the delivery of an early reward as observed in (Hollerman and Schultz, 1998) (because once the representation of the stimulus is cancelled by the reward occurrence the prediction of pause at the usual time of reward delivery naturally disappeared). This reset device made the model also able to cope with of the timing variability (for example it can account for the results observed in Fiorillo et al., 2008). Nevertheless, exactly as the model of (Montague et al., 1996), the model presented by (Suri and Schultz, 1998, 1999) supposed a infinitely precise timing mechanism (because the weight update produced learning only in representational element that had just shut off), and therefore it completely disregarded all the involvement of dopamine activation due to the noisy measurement of the temporal intervals.

1.3.2 TD models and the representation of time

The way time is represented in RL models of the dopamine system raised increased interest during the last decade. Here two attempts to fix part of the mismatches between theory and data will be presented.

A solution to the limitations of the CSC was suggested by (Ludvig et al., 2008). They introduced a more realistic temporal stimulus representation for the TD model called "microstimulus" representation that constituted a temporally smeared version of the CSC (see Figure 1.5). Aside from the detailed way this representation was constructed, two important features of the new representation deserve attention. First, in the the microstimulus representation each feature activates and, thus is eligible for learning, for a temporal range (and not for a single time step as in the CSC). Second, the temporal precision of the features decreases with time. This last feature makes the model naturally able

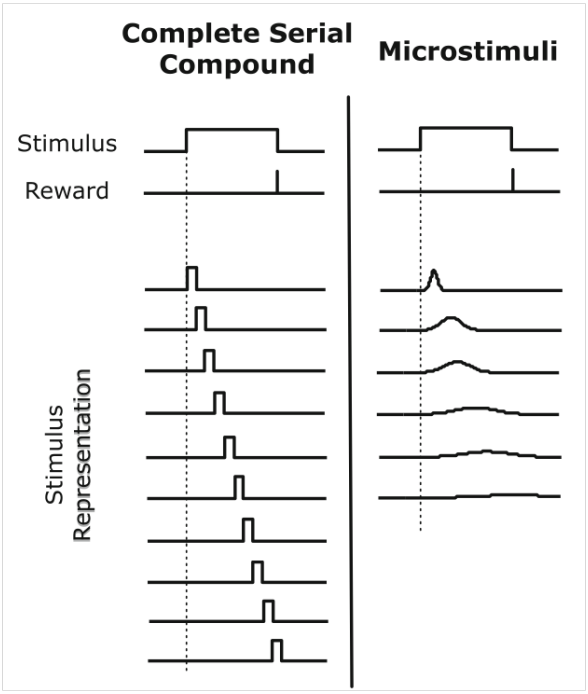


Figure 1.5: The CSC representation (left) and the microstimuli representation (right).

to cope with timing noise. In addition to this smeared temporal representation in the microstimulus model the reward is treated as a detectable event and therefore associated with its own temporal representation. The last (minor) change of the model consists in using the TD(λ) backup for prediction learning (specifically in the accumulation eligibility trace version). The predictions arising from this new TD machinery fitted the previous data considerably better than previous models. The model responded to reward omission with a shallow, extended negative TD error starting around the time the reward was expected. It is also able to behave similarly to the data reported in (Hollerman and Schultz, 1998) when an early reward is delivered (because the microstimuli representing the reward itself cancelled the positive prediction generated by the cue). More interestingly the miscrostimulus model can account for more recent data (Fiorillo et al., 2008 ;but see also Kobayashi and Schultz, 2008 for similar results). In that study monkeys were trained on a classical conditioning task in which the five different cues were associated with five different ISI's spanning from 1 to 16 s. As expected as a consequence of temporal discounting the dopamine response to the cue decreased with the ISI. Surprisingly they found that the dopamine response to the reward increased with the cue-reward interval resulting similar to the response to a completely unpredicted reward for the largest ISI. This last result was interpreted as resulting from a temporal precision of the neural expectation that sharply

declined with the interval and is compatible with the coarse temporal representation of the microstimulus model (but not with the perfect clock implicitly assumed in the CSC; see Gershman et al., 2014). Although some predictions of the model proposed in (Ludvig et al., 2008) are strictly connected with the introduction of a temporal representation for the US, the mayor improvements generated as a consequence of the coarse representation of the passage of time naturally associated with the microstimuli .

A completely different approach to the problem of timing in theories of the dopamine system was suggested in (Daw et al., 2006). They proposed a partially observable semi-Markov TD model in which dopaminergic prediction errors reflect inference over the hidden relevant variables describing the system. As mentioned at the beginning of this chapter the notion of POMDP refers to the fact that the underlying state is unknown to the agent and can only be inferred from sensory data. In addition in the semi-Markov framework the agent is also uncertain about the amount of time that has elapsed since entry into the current state and all the transition probabilities depend also on the precise dwell time distribution corresponding to each state of the system. A more detailed analysis of this alternative approach is beyond the purpose of this thesis. However, it is interesting to point out that the standpoint proposed in (Daw et al., 2006) was pioneering in suggesting that the reward prediction errors signalled by dopamine neurons operated over hidden (or 'belief') states.

The belief-state model and the microstimulus model have been generally seen as competing alternatives. The latter approach, or more in general an approach relying on a distributed element temporal representation, has been long considered preferable for several reasons (Gershman et al., 2014). From a pure computational standpoint this approach naturally fit with standard RL models of the basal ganglia, which mostly use a linear function approximation architecture. At neural level recent data (Adler et al., 2012; Mello et al., 2015) have provided direct evidence for a representation of time in the striatum that is distributed over a set of neurons. more in general any approach relying on a distributed element temporal representation can be suitable to analyse how timing affect the activity of dopaminergic neurons. In the last part of the next chapter an alternative internal temporal representation (originally proposed in Shankar and Howard, 2012) will be discussed and I will study in detail interesting predictions that can be done when this representation is embedded in the machinery of TD learning.

1.3.3 TD models and decision making under uncertainty

Despite their ability to reproduce the activity of dopamine neurons in classical and instrumental conditioning TD models have been scarcely applied to the study of the dopamine system in decision making tasks. One important advance in this direction have been proposed in (Rao, 2010). In that work the author proposed a neural POMDP model and applied it to a reaction-time version of the well-known random dots motion discrimination task used to study decision making in primates (Roitman and Shadlen, 2002; but see also Shadlen and Newsome, 2001).

Briefly, the animal had to decide the direction of motion of the coherently moving dots and was rewarded for correct decisions. The coherence of the motion varied from trial to trial making sometimes the task difficult for the animal. While making the decision the animal did not have access to the property of the stimulus relevant for obtaining the reward (in this case the direction of the dots motion). For this reason the model assumed that the noisy observations provided from the environment were used to calculate a belief about the motion direction. This belief was subsequently sent to an actor-critic architecture and the TD error signal was used to model the dopamine response. To cope with the continuous nature of belief space the model implemented function approximation using radial basis function as feature vector (both for the critic and for the actor). The value function was learnt using stochastic gradient descent TD(0) methods. The policy was represented using a softmax distribution and trained with a standard policy gradient algorithm. The model fitted the monkeys' behaviour, correctly reproducing the performance and the reaction times observed in the data. After learning, as intuitively expected, the actor learnt to make a decision only when the belief about one of the two direction was sufficiently high and keeping accumulation evidence otherwise. More importantly the TD signal developed after learning caught all the remarkable characteristics of the dopamine firing rate as reported in (Nomoto et al., 2010). In particular the TD signal showed a two component response: at the beginning of the dot motion the TD error was in all trials, and increased later only for sufficiently high value of coherence motion. A graded response to the reward delivery similar to that observed in the recordings was also reproduced by the TD signal.

These results strongly suggested the existence of neural underpinnings of RL that go beyond simple paradigms of classical and instrumental conditioning, and opened interesting perspectives for a theoretical analysis of decision making in the brain.

1.4 Neural substrate of RL in the brain

As it has been mentioned above the correspondence between RL and neural mechanisms for decision making in the brain is deeper than the simple parallel between TD learning and the phasic bursts of dopamine neurons.

The idea that an algorithm similar to the actor-critic could be implemented in the basal ganglia has been confirmed by many electrophysiology and imaging studies (see Joel et al., 2002 for a review on anatomical perspectives). According to the current view of RL in the basal ganglia, the ventral striatum is generally associated with prediction learning whereas the dorsal striatum is assumed to subserve action selection and policy learning.

The ventral striatum is a good candidate to act as the critic of the formal theory for a variety of reasons: it projects to, and receives projections from, the dopaminergic system (Joel and Weiner, 2000); it shows sustained activity during the period in which rewards are expected (Schultz et al., 1992; Setlow et al., 2003). In addition the ventral striatum, unlike other portions of the striatum, is connected to dopaminergic neurons that project to all regions of the striatum (Joel and Weiner, 2000). This last property is fundamental for generating a dopamine signal that acts as the prediction error in the actor-critic architecture and guide both prediction learning and action selection. Other brain areas fulfilling similar requirements and closely interconnected anatomically with the ventral striatum are the orbitofrontal cortex and the amygdala. Indeed, it is likely that these three areas may work together to implement the critic.

The association of the the actor with the dorsolateral striatum is supported by the implication of this part of the basal ganglia in habitual behaviour (Packard and Knowlton, 2002; Daw et al., 2005; Wickens et al., 2007). Furthermore many results provided additional evidence for the idea that the critic is related to the ventral striatum and the actor is related to the dorsal striatum. O'Doherty et al., 2004 conducted an fMRI study in human encountering that BOLD activation in the dorsal striatum was related with prediction errors only when active choice behaviour was required (whereas the activation in the ventral striatum correlated with prediction errors also in pavlovian contingencies). An electrophysiological study in rats (Daw, 2003) showed that neurons in the dorsal striatum represented actions whereas neurons in the ventral striatum represented predicted rewards.

Although some studies suggested that dopamine response could reflect different RL algorithms (see Morris et al., 2006; Roesch et al., 2007), the role of the striatum in reward-

related learning is ubiquitous. Correlates of prediction errors in the dorsal and ventral striatum have now been seen in multiple studies. A study reported in (Schönberg et al., 2007) showed that the ability to learn optimally was highly correlated with striatal BOLD activation reflecting reward prediction error.

To summarize: during the last years TD models and the reward prediction error hypothesis have gone beyond the simple parallel between the phasic dopaminergic signals and the TD learning algorithm. Indeed they have further linked algorithmic ideas from RL to possible underlying neural substrates, specifically, to learning and action selection in the basal ganglia mediated by corticostriatal dopamine-dependent plasticity. Converging evidence from a wide variety of recording and imaging methods supports this hypothesis.

Chapter 2

Dopamine, temporal expectations, and TD models

Keeping track of time is fundamental to learn about the sometimes-delayed consequences of actions and to guide optimal rewarding behaviour. Although timing and prediction error driven learning have historically been treated as independent processes, growing evidence seems to indicate that they are strictly connected.

Distributed sets of brain areas, especially the cortico-basal ganglia circuits, known for being involved in reinforcement learning and decision making (Lau and Glimcher, 2008; Cai et al., 2011; Lee et al., 2012; Ding and Gold, 2013; Lee et al., 2015), have been implicated in the representation of time across temporal intervals (Hinton and Meck, 2004; Meck, 2006; Wencil et al., 2010; Merchant et al., 2013).

Recordings from the striatum (of both rats and monkeys) have indicated the existence of cluster of neurons which activity encoded time in a way that resembled the one required by RL models (Jin et al., 2009; Adler et al., 2012; Mello et al., 2015).

The information carried by these neurons appeared to be sufficient to decode time from the population response (Jin et al., 2009), and the time estimates decoded from the population activity predicted the animal timing behaviour (Mello et al., 2015).

In addition a deep connection between the dopamine reward prediction errors and the processing of time has been demonstrated in a wealth of studies. The phasic activation of dopamine neurons crucially depended on relative timing of the relevant (reward-related) task events (Fiorillo et al., 2008; Nomoto et al., 2010). Negative tonic modulations of the dopaminergic neurons basal firing rate seemed to reflect temporal expectations (Bromberg-Martin et al., 2010; Pasquereau and Turner, 2015). Data indicated that the temporal precision of reward (or reward related) predictions in dopamine neurons sharply

declined with time (Fiorillo et al., 2008).

These studies have suggested that a distribute elements temporal representation which precision decreases with the interval duration could be appropriate to represent time in RL models of the basal ganglia (Gershman et al., 2014). However a few different temporal representations fulfil these requirements.

This chapter analyses reward prediction errors deriving from a precise choice of temporal representation and compares this prediction with current available recordings. The next section will briefly review how the activity of dopamine neurons relates with temporal expectations. Then I will introduce in the TD machinery a scale invariant representation of time (Shankar and Howard, 2012, 2013) and I will compare the predictions generated from the model with available data. This study will determine the choice of the temporal representation adopted in chapter 3 and in chapter 4.

2.1 Dopamine recordings and temporal expectations

Pioneering observations showing that the dopamine prediction error signals were sensitive to temporal expectation of reward and predictors were reported in (Schultz et al., 1993; Hollerman and Schultz, 1998), as discussed in chapter 1. The general findings of these early studies were that when the relative timings of relevant task events was deterministic, dopamine neurons fired only to the first stimulus. Besides, dopamine neurons responded to events whose timing was variable, provided that the variability in event timing exceeded some threshold (200-500 ms; Daw, 2003)

Fiorillo et al., 2008 performed three different experiments to shed light on the way how temporal expectations affected the dopamine reward prediction errors. They found that dopamine responses were sensitive to temporal aspects of reward expectation, both at the time of the reward-predicting stimulus and at the time of the reward (see Kobayashi and Schultz, 2008 for similar results). In addition, the dopamine response was not highly precise and the precision declined sharply with the interval duration.

In the first Pavlovian task monkeys were trained to expect the reward after a fixed delay interval of variable duration. The ISI varied from 1-16 s depending of the visual stimulus presented at the beginning of each trial. The response to the CS was found to decrease with the duration of the ISI, as expected taking into account temporal discounting (see Figure 2.1a, left). However, in contrast with previous observations with short fixed intervals (typically between 1 and 2 seconds), when the duration of the ISI increased the reward caused a substantial activation of dopamine neurons (see Figure 2.1a, right).

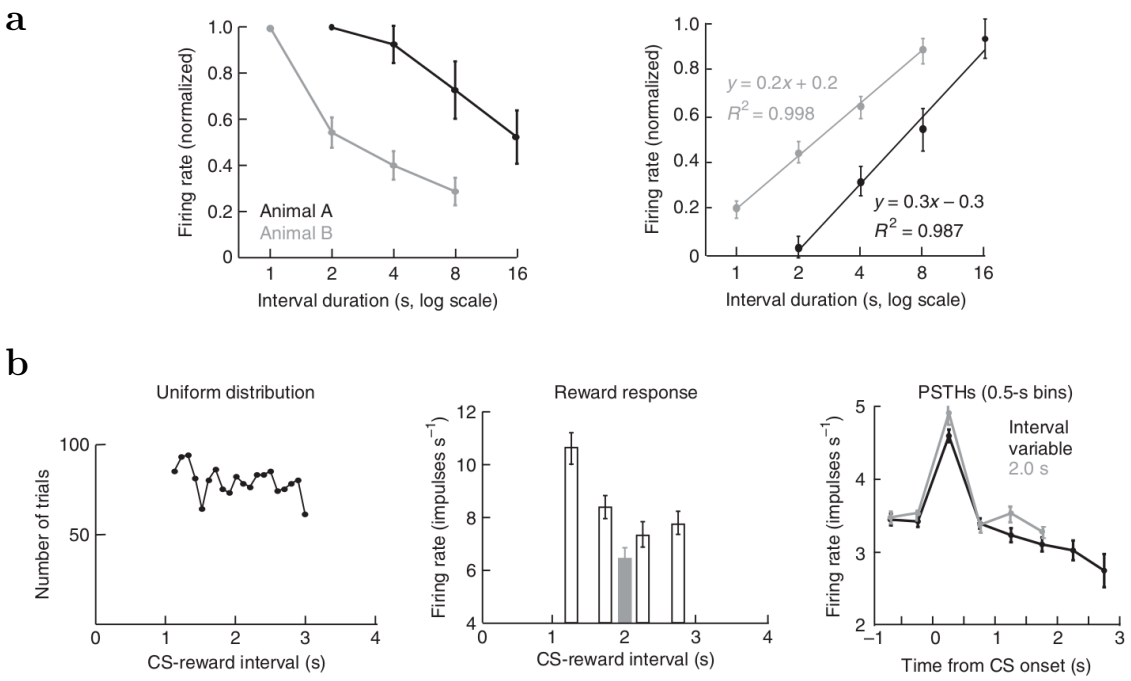


Figure 2.1: **Dopamine responses and temporal expectation.** (**a**) When the reward was delivered after different ISI of fixed duration the response of dopamine neurons to the CS decreased with the ISI duration (left). The response to the reward showed an opposite trend (right). (**b**) When the ISI duration varied according to a uniform distribution (left) the dopamine activation to the reward decreased with the ISI. The gray bar represents the response to the reward delivered after a fixed interval of 2 s (middle). In the variable interval following the CS the dopamine tonic activity showed a decreasing trend (right, black line). Note how for a 2 s fixed ISI the activity before the reward delivery was suppressed with respect to its baseline calculated before the cue presentation (right, gray line). Adapted from (Fiorillo et al., 2008).

The authors concluded that the momentary expectation of reward decreased with interval duration (likely accordingly to Weber's law), and that neurons showed a significant response for long intervals because of the poorly precise reward prediction.

The fact that temporal expectation at neural level was not very accurate was confirmed by another experiment reported in the same work. The activation of dopamine neurons seemed to indicate that the neural expectation was relatively strong even after just half of the usual ISI. Indeed, the dopamine response to the reward was greater than the activation at the usual time of the reward delivery, but much less than the activation to an 'unpredicted' reward. In addition the firing rate showed a slight decrease before

the usual time of reward delivery with respect to its baseline calculated before the cue presentation (see the gray line in Figure 2.1b, right). This effect can be interpreted in terms of the animals' timing noise as a form of negative prediction error due to a weak expectation of reward even in a short period before the end of the usual ISI. Finally Fiorillo and colleagues analysed the dopamine response to a reward delivered after a variable ISI (uniformly chosen between 1 and 3 seconds). They found that neuron responded with greater activation to the reward after short ISI, consistently with the idea that the reward expectation grew with the elapsed time. However neurons were only slightly more activated by reward delivered after the variable interval than after a fixed 2-s interval (again consistently with a low precise neural temporal expectation; see Figure 2.1b, center, gray bar). In addition the firing rate of dopamine neurons gradually declined as the stimulus-reward interval elapsed in the absence of reward (Figure 2.1b, center, black bars). This was interpreted as a form of negative prediction errors connected with the reward expectation.

Similar patterns of activity were observed in several studies that have analysed the neural temporal expectation related to events different from primary rewards (Bromberg-Martin et al., 2010; Nomoto et al., 2010; Pasquereau and Turner, 2015). These works reported that, when the interval preceding the event was variable, dopamine neurons carried two distinct signals. On the one hand a tonic decreasing activation anticipated the time of the upcoming event. On the other hand the occurrence of such an event produced a phasic dopamine burst that depended on the elapsed interval.

All these studies suggested that dopamine neurons encoded the timing distribution of upcoming task events through gradual decreasing modulation in their tonic activity. Figure 2.2a (right) shows how the tonic dopamine activity anticipated the start cue occurrence for three different ITIs, all of variable duration. At each moment when the timed event failed to appear but could have potentially occurred the pattern of activity resembled negative reward prediction errors related with temporal expectations. This tonic signal appeared to be modulated by the timed event hazard rate, i.e the probability of the event to occur given that it has not occurred yet. The first panel in Figure 2.1b (left) shows the anticipatory activity for a fixed ITI of duration equal to 2.2 seconds. Note the decreasing tonic modulation slightly before the cue appearance, and the similarity with the result reported in (Fiorillo et al., 2008) (shown in Figure 2.1b, right, gray line).

Contrasting event-evoked dopamine responses were instead reported in these studies. The majority of the data indicated that the evoked responses were smaller for longer intervals. Fiorillo et al., 2008 found that the dopamine neurons were less activated by

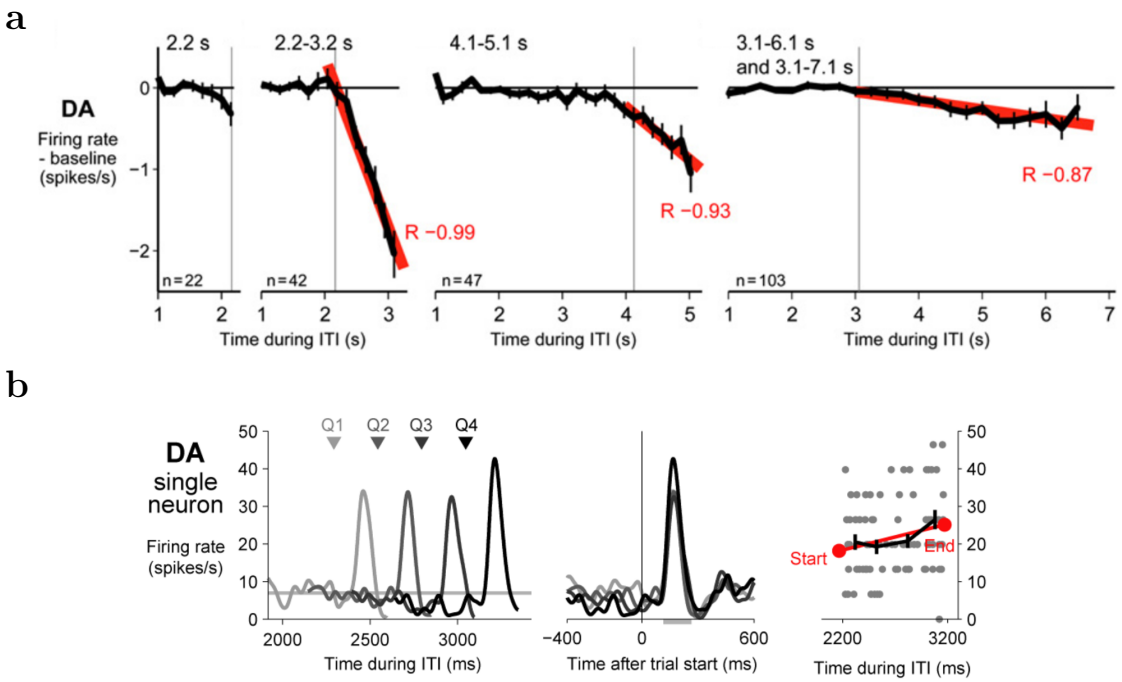


Figure 2.2: **Dopamine response after variable ITIs.** (**a**) Anticipation dopamine activity preceding the star cue presentation for ITIs of different duration. Note the decreasing tonic modulation slightly before the cue presentation for the fixed duration ITI (first panel on the left). (**b**) Cue-evoked response sorted according to the ITI duration, for an ITI that lasted between 2.2 and 3.2 seconds (flat distribution). Adapted from (Bromberg-Martin et al., 2010).

the reward for longer ISIs (Figure 2.1b, center, black bars). A similar pattern of evoked response was reported in (Pasquereau and Turner, 2015), that investigated how dopaminergic neurons changed their activity in relation to a go signal (triggering a reaching movement) that varied temporally across trials with a flat distribution from 0.5 to 1.5 s. And also in (Nomoto et al., 2010) in relation to the activity following the visual appearance of random moving dots, that occurred between 0.5 and 1.5 seconds after a previous cue. However (Bromberg-Martin et al., 2010) reported an opposite trend: dopamine responses to a start cue were larger for longer ITIs (see Figure 2.2b).

In what follows I will analyse the reward prediction error deriving from a specific TD model. Neither the learning algorithm nor the temporal representation I will adopt are new. The TD algorithm is the one used in (Montague et al., 1996) with an additional reset mechanism (similar tothe one suggested in Suri and Schultz, 1999). The temporal representation has been developed in (Shankar and Howard, 2012, 2013). The analysis will suggest that the apparently contrasting evoked responses are due to the durations of

the variable interval being timed, namely between 0.5 ans 1.5 seconds (so relatively short) in (Nomoto et al., 2010; Pasquereau and Turner, 2015) and between 2.2 and 3.2 seconds (or longer) in (Bromberg-Martin et al., 2010). The results of the simulation will show that an adequate choice of the temporal representation can enable the TD model to cope with all the current available data.

2.2 The TD model

2.2.1 Temporal representations

The onset of each reward-predictive stimulus initiates a vector of sub-states $\mathbf{x}(t) = \{x_1(t), x_2(t), ...\}$ that tracks the passage of time. From a computational perspective this sequence of states corresponds to the stimulus features for value computation. Two different temporal representations will be adopted in the simulations, and are briefly discussed below. Each of them makes different assumptions about the temporal precision with which animals track the passage of time.

Complete serial compound The 'complete serial compound' (CSC) is the temporal representation used in the original application of TD learning to the dopamine system (Montague et al., 1996; Schultz et al., 1997). At each time step after the stimulus presentation, only one state of the vector representation is active. This means that $x_i(t) = 1$ only i time steps after the stimulus onset, and it remains equal to zero otherwise (see Figure 1.5, left). The representation completely excludes temporal generalization between consecutive time steps and therefore it makes the implicit assumption of a perfect clock.

Shankar-Howard representation The representation of time proposed in (Shankar and Howard, 2012, 2013; hereafter referred to as the SH representation) belongs to the class of distributed elements representations that allows generalization between nearby time points (similar to the one adopted in Ludvig et al., 2008).

The reward-predictive stimulus (a pulse of duration equal to one time step dt) is represented at different latencies (or nodes) τ_m from its onset through a set of function which precision decreases with the passage of time (see Figure 2.3). Indicating with t_{CS} the onset time of the stimulus an explicit mathematical realization of the SH representation is the following:

$$x_m(t) = \frac{1}{|\tau_m|} C(k) \int_{t_{CS}-t+dt}^{t_{CS}-t} \left(\frac{\tau'}{\tau_m}\right)^k e^{-k\frac{\tau'}{\tau_m}} d\tau' \tag{2.1}$$

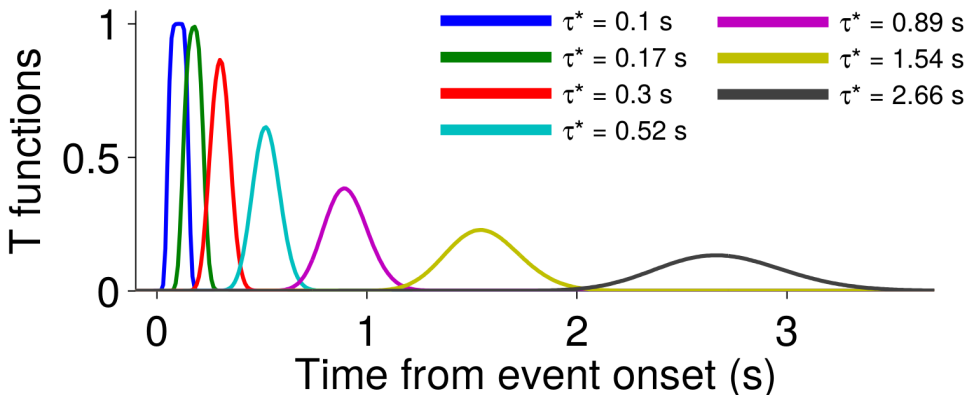


Figure 2.3: The SH representation with nodes distributed uniformly on a logarithmic time scale.

where $C(k) = k^{k+1}/k!$ and the parameter k controls the smear in the representation (the larger is k the more accurate is the representation). The distribution of the nodes crucially affects the properties of the SH representation. Accordingly to the optimality principles described in Shankar and Howard (2013) the nodes are chosen to be distributed uniformly on a logarithmic time scale:

$$\tau_{min}, \tau_{min}(1+c), \tau_{min}(1+c)^2, \cdots \tau_{min}(1+c)^{(N_{max}-1)} = \tau_{max}. \tag{2.2}$$

where N_{max} is the number of nodes (and therefore the number of feature vectors) used by the SH representation. The parameters are chosen accordingly to the shortest and the longest temporal interval to be timed. Here I will use $\tau_{min} = dt = 0.1$ s and $\tau_{max} = 10$ s. To minimize both information redundancy and information loss, for a given value of k, the parameter c will be picked in order to approximately satisfy the following relationship (Shankar and Howard, 2013):

$$\frac{k}{(k-1)\sqrt{(k-2)}} \approx \frac{c}{1+c} \tag{2.3}$$

2.2.2 Learning algorithm

The model learns through the linear TD(0) algorithm (Sutton and Barto, 1998; Schultz et al., 1997; Montague et al., 1996), that has been discussed in subsection 1.1.4. For completeness the basic equations of the algorithm will be repeated below. At each time step, the estimated value is determined by a linear combination of stimulus features:

$$V(t) = \mathbf{w}^T \cdot \mathbf{x}(t) = \sum_{i=1}^{N} w_i \cdot x_i(t) \tag{2.4}$$

and N is the total number of features. The weights are updated according to the following learning rule:

$$\Delta w_i = \alpha \delta(t) x_i(t-1) \tag{2.5}$$

where α is the learning rate, $\delta(t) = r(t) + \gamma V(t) - V(t-1)$ is the usual reward prediction error, and γ is the discount factor.

2.2.3 Additional reset mechanism

A simple modification of the TD algorithm described above can be achieved by introducing a 'reset' mechanism that sets all the stimulus features to 0 after reward arrives:

$$x(t) = 0 \quad \forall\, t \geq t_r \tag{2.6}$$

where t_r indicates the time step of the reward delivery.

The introduction of the reset crucially affects the convergence of the TD algorithm. In particular in this case the value function converges to:

$$V(t) = E\left[\sum_{i=0}^{\infty} \gamma^i r(t+1+i) | r(t') = 0 \,\forall\, t' < t\right] \tag{2.7}$$

See Appendix A for a formal justification of the above equation.

The TD model without reset is fully defined by the choice of the temporal representation and by the Equation 2.4 and Equation 2.5. The complete definition of the TD model with reset requires the additional Equation 2.6. In all simulations, the discount factor and the learning rate were fixed to the values $\alpha = 0.1$, $\gamma = 0.97$, and 10 time steps were interpreted as a unit of 1 s.

The TD model yielded asymptotic results after a number of trials that depended on the presence of the reset mechanism and on the temporal representation (for small values of the parameter k the number of trial increased). The results shown in what follows represent the average of $n = 3000$ simulated trials taken after the TD model showed asymptotic properties.

2.3 Results

I used the TD model described above to simulate experiments of simple acquisition in which the reward is delivered at random times. In all the simulations a CS was presented at time step $t_{CS} = 0$ and the reward occurs between the time steps t_r^{min} and t_r^{max} with

flat probability distribution $f(t) = P(r(t))$. I analysed the TD model predictions with and without the reset mechanism, and with different temporal representations, namely for the CSC and for the SH representation with different level of temporal precision (i.e different values of the parameter k). The extremes of the flat distribution (t_r^{min} and t_r^{max}) were varied to analyse the effects of timing and temporal uncertainty. In Appendix A I will show that the TD model with the reset mechanism in the simulations of the simple acquisition described here produce a RPE at reward delivery that is equivalent (from an algorithmic point of view) to the RPE generated by the occurrence of a general task event that resets the representation of previous stimuli. This equivalence justifies the study preformed in what follows and the comparison with the available data described in section 2.1.

2.3.1 Results: TD model without reset

I simulated the TD model without reset for an experiment of simple acquisition similar to the one studied in (Fiorillo et al., 2008). In each trial a CS was presented at time 0, and reward was delivered after 1 s to 3 s (with flat distribution $f(t)$), i.e between time steps $t_r^{min} = 10$ and $t_r^{max} = 30$. Hereafter I will refer to the period of time between t_r^{min} and t_r^{max} as the possible reward window, and to the interval between the CS presentation and the beginning of the possible reward window (i.e $t < t_r^{min}$) as the pre-reward period.

The TD model with CSC and without reset produced RPEs at reward delivery that were roughly independent from the ISI duration (see Figure 2.4b). The pattern of RPEs across different ISIs (see the red line in Figure 2.4a) resembled a flipped distribution of experienced ISIs (i.e it resembled the flipped probability distribution $f(t)$). These RPEs were clearly at odd with the results encountered in (Fiorillo et al., 2008). Data from that study suggested that the activation of neurons at reward delivery were greater on trials with shorter ISIs (see Figure 2.1b, middle). Also, unlike the pattern of constant negative RPEs across different ISIs produced by the simulations, data suggested that the tonic firing rate tended to gradually decline as the variable interval elapsed (compare the red line in Figure 2.4a and, the black line in Figure 2.1b, right).

Noteworthy was the temporal evolution of the value function (see the blue line in Figure 2.4, left). The value exponentially increased from the time of the CS presentation to the first time of possible reward delivery. After the time step t_r^{min} (corresponding to an ISI of 1 s) the value started to decline and became equal to 0 at time step t_r^{max} (corresponding to the maximum ISI, i.e 3 s). The value function can be calculated analytically for the TD model with CSC and it can be proved to decreased in a way that is roughly proportional to

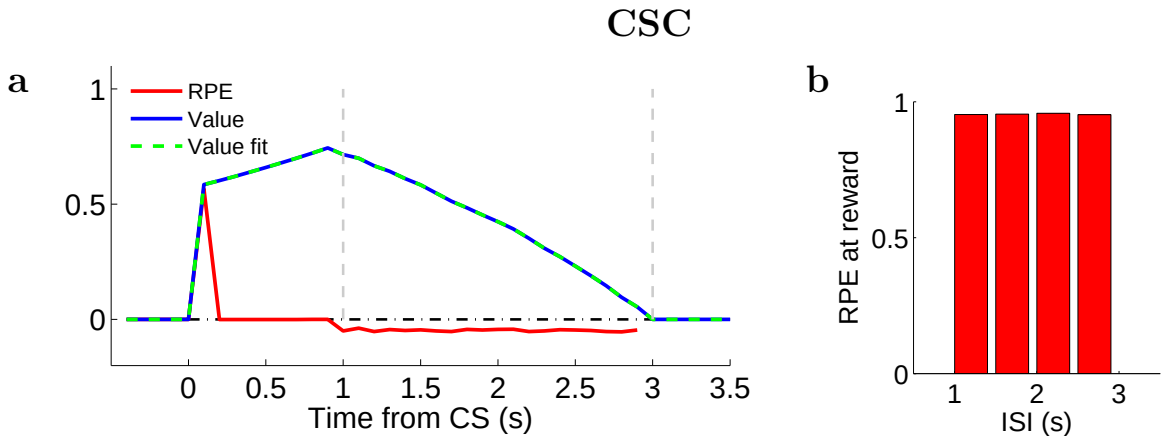


Figure 2.4: **TD model without reset and with CSC.** (**a**) TD whitout reset and with CSC produced a pattern of negative de RPEs (red line) within the possible reward window (marked by the dashed gray lines) that resembled a flipped probability distribution. The simulated value function (blue line) decreased as the ISI duration increased and coincided with its analytic expected value represented by green dashed line. (**b**) The RPEs at reward delivery were roughly independent from the ISI duration.

the increase of the cumulative distribution function $F(t) = \sum_{i=0}^{t-1} f(i)$ (see Appendix A). This temporal evolution reflected the definition of the value function, that represented the expected discounted sum of future rewards and that decreased as time elapsed during the possible reward window. However the value temporal evolution obtained from the TD model without reset was clearly at odd with the intuitive idea of reward expectation. Given that the reward has not occurred yet, its expectation should indeed increase during the possible reward interval, becoming bigger and bigger for longer ISIs.

Introducing the SH representation in the TD model without reset did not help to fix these problems (see Figure 2.5). Independently from the precision of the representation the value function converged roughly to the same temporal profile encountered with the CSC, and thus the TD model without reset and with the SH representation produced similar mismatches between simulations and data.

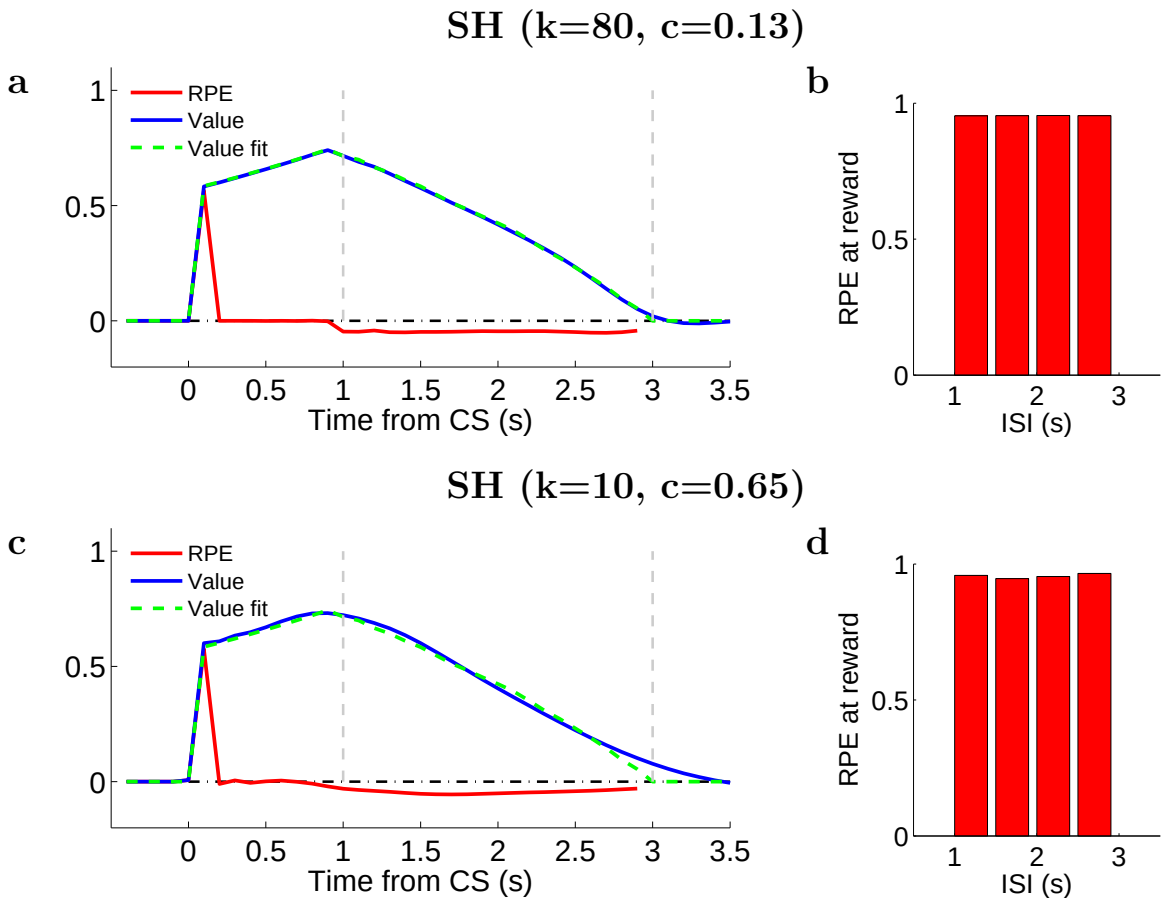


Figure 2.5: **TD model without reset and with SH representation.** Same that in Figure 2.4 but using two differently precise SH representations in the TD model without reset. (**a,b**) For a highly precise representation ($k = 80$) the results were almost identical to those obtained with the CSC. (**c,d**) Although the precision of the representation considerably decreased ($k = 10$) the TD model showed similar asymptotic behaviour.

2.3.2 Results: TD model with reset

The introduction of the reset mechanism rectified key inconsistencies between the data reported in (Fiorillo et al., 2008) and the simple CSC TD model. The RPEs at reward delivery showed a clear tendency to decrease for long ISIs, behaving similarly to the response of neurons (compare Figure 2.6b with the middle panel in Figure 2.1b). The pattern of RPEs across different ISIs (red line in Figure 2.6a) showed only a slight negative modulation during the possible reward window, unlike the clear decreasing profile with elapsed time in Figure 2.1b (black line on the left).

Importantly, unlike the TD model without reset, the the temporal profile of the value function (blue line in Figure 2.6a) was clearly consistent with the idea of reward expectation: the value increased as the time elapsed during the possible reward window, resembling the fact that the reward became more and more expected given that it has not occurred yet. The profile of the value function can be calculated analytically for the TD model with reset. This analytic profile coincided with the simulated value function of the TD model with CSC, and in each time step converged to the expected discounted sum of future rewards conditioned to the fact that the reward has not occurred previously (see Appendix A). Assuming a perfect internal clock (as implicit in the CSC), at the end of the possible reward windows the occurrence of the reward became fully predicted and the value function approached to one.

Using the SH representation to track the passage of time, for the same variable ISI (between 1 s and 3 s), the results crucially depended on the precision of the representation (determined by the parameter k). The value function deviated from its analytic profile during the possible reward window (Figure 2.7a,c,e). The time step in which the deviation began and the magnitude of the deviation depended on the precision of the representation. When the accuracy was very low ($k = 10$) the deviation was remarkable even for ISI of medium duration. The RPEs at reward delivery decrease with the ISI for a high precise representation (Figure 2.7b) but exhibited an opposite trend when the precision sharply decreases (Figure 2.7f). For a representation of average precision the decreasing trend of the RPEs was reversed for very long ISIs (Figure 2.7d; note the similarity with the pattern of responses reported in (Fiorillo et al., 2008)). To summarize: the results with the SH representation matched those obtained with the CSC at different levels of accuracy (that depended on the parametr k). The value function tracked its analytic profile that corresponds to reward expectation conditioned to the fact that the reward has not occurred. However the TD model, due to the imperfect timing mechanism associated with the SH representation, was unable to produce a perfect tracking of the analytic value

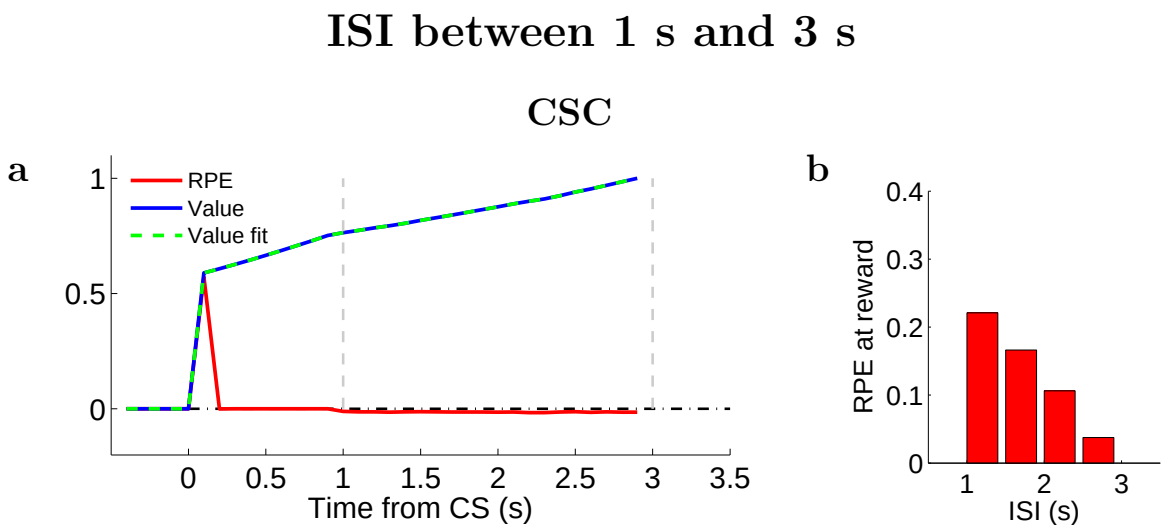


Figure 2.6: **TD model with reset and CSC for ISI between 1 s and 3 s.** (**a**) TD whit the additional reset mechanism and with CSC produced a value function (blue line) that increased with the ISI duration, consistently with the intuitive idea of reward expectation. The analytic expected value (green dashed line) again coincided with the simulated value function when the temporal representation was the CSC. (**b**) The RPEs at reward delivery decreased with the ISI duration.

profile for long ISIs. As time elapsed during the possible reward window, its tracking cease to be sufficiently accurate. As a consequence the momentary expectation of reward decreased, producing large RPEs at reward delivery for long ISIs (this effect is particularly evident for the SH representation with $k = 10$ and $k = 40$, depicted in Figure 2.7d,f).

I simulated the TD model with reset for different durations of the pre-reward period and for different durations of the possible simulation window. The ISIs duration were chosen to allow a comparison with current available data. Although the majority of the studies described in section 2.1 analysed the response of dopamine neurons to events which timing was variable different from primary rewards, in Appendix A I show that, concerning to the dependence on temporal expectation, the TD model with reset would produce similar results independently on the nature of the resetting event. The algorithmic equivalence reported in Appendix A justifies the approach employed in what follows.

The TD model with reset and with CSC produced similar results independently from the temporal parameters that determined the ISI duration (i.e the pre-reward and possible reward window duration). The value function always tracked its analytic profile and the RPEs after reward occurrence decreased for long ISIs.

The model simulated with the SH functions of high precision closely resembled the value profile and the RPEs obtained with the CSC when the pre-reward period was short (Figure 2.9a-d). However for an ISI variable between 5 s and 7 s the accuracy in reproducing the analytic value profile dropped toward the end of the possible reward window (Figure 2.9e) and the RPEs at reward delivery increased for very long ISIs (Figure 2.9f). A very low precise SH representation (see Figure 2.11) for all the simulated ISI produced a pattern of RPEs that increased for long ISIs. More interesting was the situation when an average precise SH representation was incorporated in the TD model. For a short pre-reward period the tracking of time was quite accurate and the reward expectation increased during the possible reward window suppressing the RPEs at reward occurrence for long ISIs (see Figure 2.10b). This result was in agreement with the data reported in (Pasquereau and Turner, 2015). As far as the pre-reward period increased the RPEs toward the end of the possible reward window became more pronounced because of the dropping in the timing accuracy (Figure 2.10f). Note in particular that for a temporal variability similar to the one analysed in (Bromberg-Martin et al., 2010) the average precise SH representation produced a pronounced RPE at the end of the possible reward window that resemble the response of the dopamine neurons (compare Figure 2.2b with Figure 2.10d).

To summarize: the results obtained with the TD model and reset were consistent with the idea that when the reward could have been delivered but failed to occur its expectation increased. The perfect clock imposed by the CSC produced results exactly in line with this idea of increasing temporal expectation. For the SH representation the results crucially depended on the precision of the timing mechanism (determined by the parameter k), and on the properties of the interval to be timed (namely the pre-reward period and the possible reward window duration). The simulation showed that the TD model with reset and with an average precise SH representation produced RPEs compatible with the responses of dopamine neuron as reported in available data.

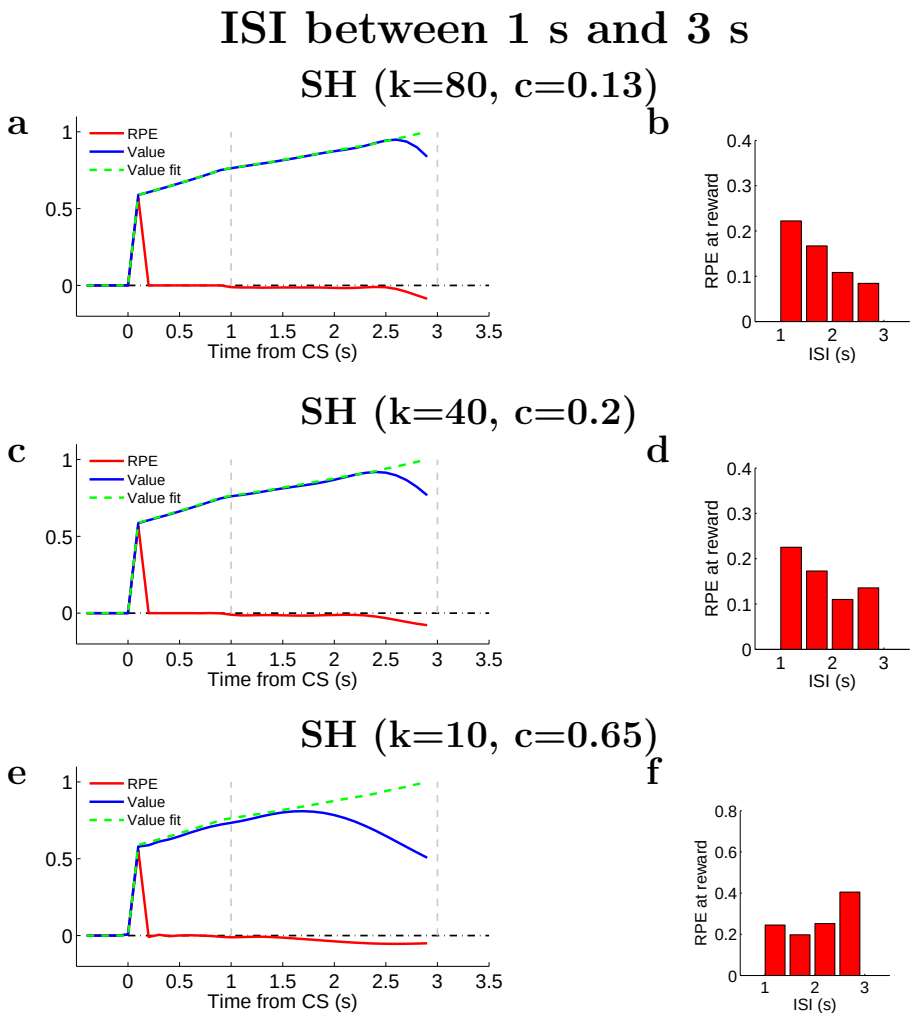


Figure 2.7: **TD model with reset and with SH representation for ISI between 1 s and 3 s.** (**a**,**b**) For a highly precise SH representation ($k = 80$) the results were quite similar to those obtained with the CSC. The simulated value function (see the **a** panel) departed from the analytic expected value for long ISIs, and this produced large negative modulation of the RPEs (red line) toward the end of the possible reward window. As in the simulation with CSC the RPEs at reward delivery (depicted in **b**) decreased for longer ISIs. (**c**,**d**) The TD with a less accurate temporal representation ($k = 40$) produced a value function that mismatched its analytic expected value for a longer period toward the end of the possible reward windows. This effect was responsible for the a slight increase of the RPEs when the reward occurred at the end of the possible reward window (see the last bin of the histogram in **d**). (**e**,**f**) When the precision of the representation further decreased ($k = 10$) the results of the simulations considerably deviated from those obtained with the CSC. Note in particular that the RPEs increased for long ISIs as shown in **f**.

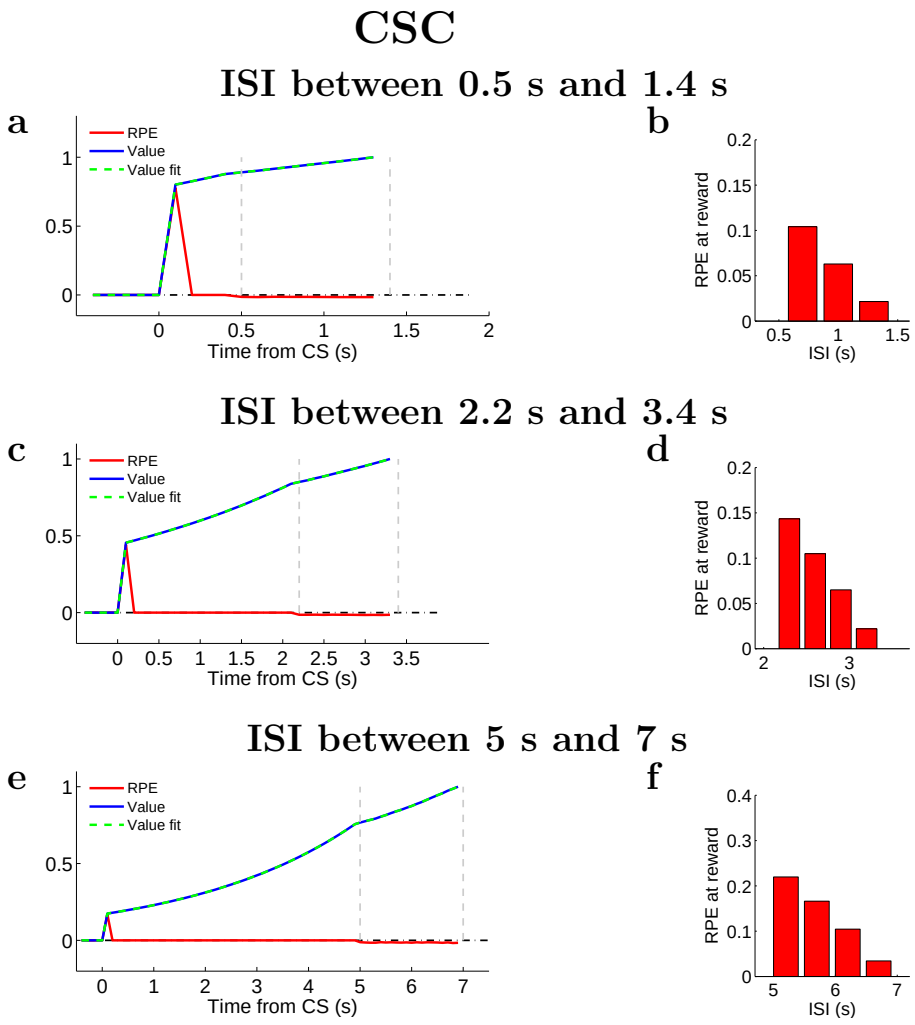


Figure 2.8: **TD model with reset and with CSC for ISI of different duration.** Independently from the duration of the possible reward window and from the value of the shortest ISI the CSC produced qualitatively similar results. The simulated value function always matched its analytic expected value (see panels **a,c,e**) and the RPEs decreased for longer ISIs (see panels **b,d,f**).

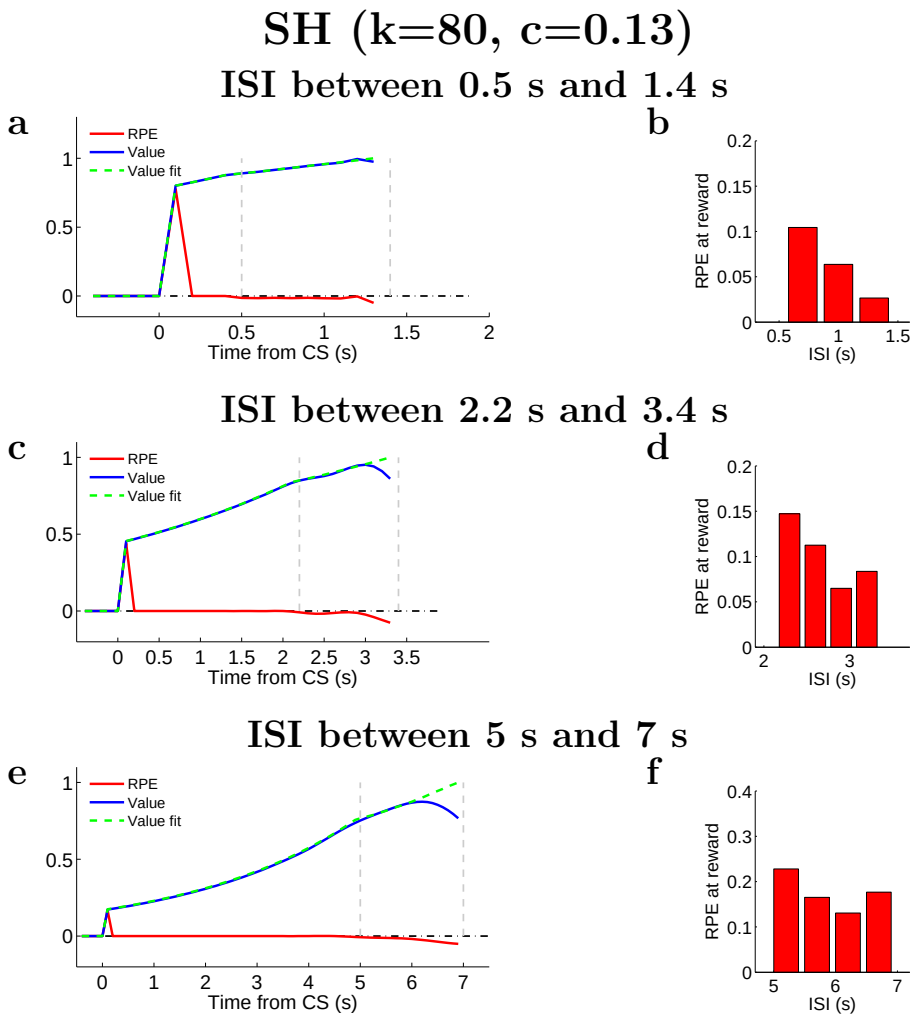


Figure 2.9: **TD model with reset and SH (k=80) for ISI of different duration.** (**a**) When the duration of the (variable) ISI was rather short (it varied between 0.5 s and 1.5 s) the highly precise SH representation ($k = 80$) approximated quite well the results obtained with the CSC. (**b**) As a consequence the RPEs at reward decreased with the ISI duration. (**c**) If the pre-reward period increased the value function (blue line) started to separate from the analytic temporal profile toward the end of the possible reward window, and the negative modulation of the RPEs became more pronounced (red line). (**d**) The RPEs at reward delivery decreased with the ISI duration, although for very long ISIs the RPEs at reward showed a tendency to increase (see the last bin of the histogram). (**e**) Both the value function and the pattern of RPEs showed a tendency similar to that of panel **c**, with long ISI worse represented. (**f**) The RPEs at reward delivery showed a decreasing/increasing pattern similar to the one described in **d**.

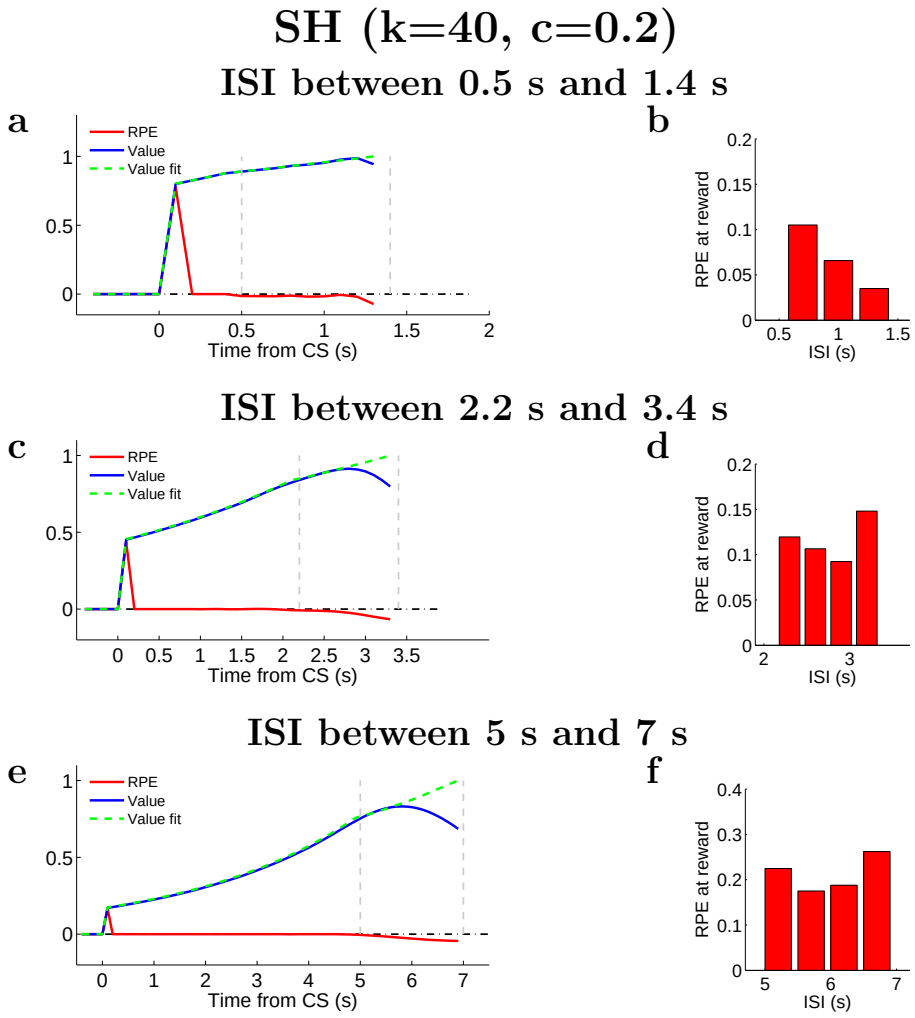


Figure 2.10: **TD model with reset and SH (k=40) for ISI of different duration.** (**a,b**) For an averagely precise SH representation ($k = 40$) and short pre-reward period results are similar to those obtained with the CSC (apart from a slight mismatching of the value function toward the end of the possible reward window). (**c,d**) However when the pre-reward period increased to 2.2 s the value function remarkably deviated from the analytic profile (see the **c** panel), and the RPEs at reward considerably increased for long ISIs (last bin of the histogram in panel **d**). (**e,f**) Results deviated more and more from those obtained with the CSC and the RPEs at reward delivery started to show an increasing tendency even for trials that lasted slightly more than the average (see the last two bin of the histogram in panel **f**).

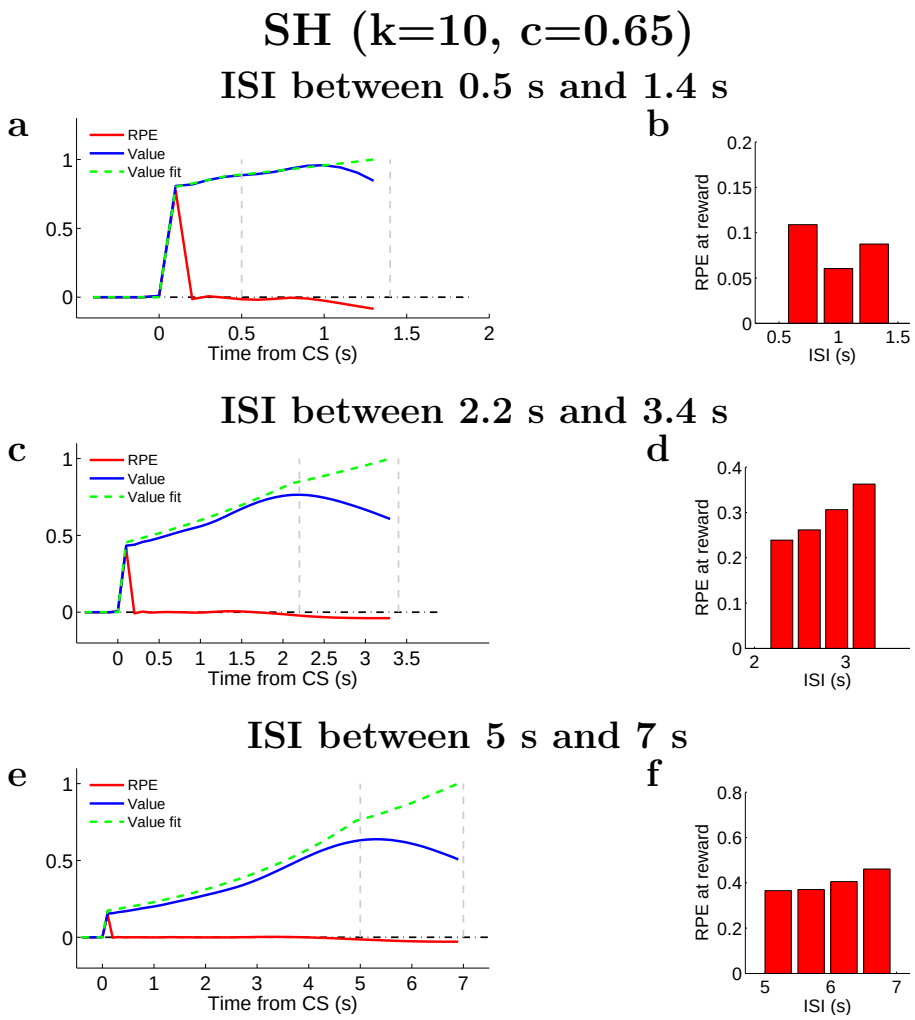


Figure 2.11: **TD model with reset and SH (k=10) for ISI of different duration.** (**a**) For a very inaccurate SH representation ($k = 10$) the results deviate from those obtained with the CSC even for a very short pre-stimulus reward. (**b**) The RPEs showed a decreasing/increasing tendency when the ISI varied between 0.5 s and 1.5 s. (**c,e**) The value function deviated from its analytic profile before the beginning of the possible reward window. This produced a slight negative modulation of the RPEs toward the end of the pre-reward period (red line). (**d,f**) The very low accuracy of the SH representation produces RPEs at reward delivery that increased with the ISI duration.

2.4 Discussion

Timing is pivotal to guide optimal rewarding behaviour. The RPE signals conveyed by midbrain dopamine neurons, which are thought to drive associative learning, have been demonstrated to be strictly related with the temporal expectation of reward and reward-related events. In particular, when the relative timing of reward-related events varied according to some predefined probability distribution, the dopamine signal seemed to reflect the probability of the event to occur given that it has not yet occurred.

Here I focused on the ability of a simple TD model to cope with current available data, simulating experiments of acquisition in which rewards were delivered at variable times according to some predefined probabilistic schedule. I analysed how the model predictions depended on two factors: the way of representing time (namely the temporal representation adopted by the model), and the presence of a reset mechanism that stopped the interval counts after the arrival of the reward.

In absence of the reset the TD model, independently from the temporal representation, gave rise to RPEs that were at odd with existent experimental findings (Fiorillo et al., 2008). In particular, the simulated value function deviated from the intuitive idea of reward expectation that would be expected to increase as time elapsed and the reward failed to occur.

The TD model with reset and CSC could explain the data reported in (Fiorillo et al., 2008). The results generated by the model were consistent with the fact that the expectation of reward increased as the reward failed to appear during the possible reward window, and its occurrence became more and more predictable. In particular the RPEs at reward delivery showed a decreasing trend for rewards delivered after long ISIs. This dependence of the RPEs on the conditioned probability of the reward occurrence is consistent with a previous study that analysed the responses of dopamine neurons in a somehow different context (Nakahara et al., 2004). In that work the authors noted that the dopamine neurons responses were related to history effects, and that the dopamine conveyed a prediction error signal based on the trial-to-trial conditional probability of reward.

Returning on the prediction of the TD model with reset and CSC, the simulation showed that, due to the implicit assumption of a perfect clock, the use of the CSC representation produced RPEs that were independent on the duration of the pre-reward period and of the possible reward window. Thus, the TD with reset and with CSC could explain data reported in (Bromberg-Martin et al., 2010).

The introduction of the SH generated results that depended on the temporal properties

of the ISI distribution to be timed, and that were crucially affected by the precision of the interval timing mechanism (encoded by the parameter k). Results were consistent with the idea that, although the reward became more expected with the elapsing of time, for long ISI the momentary expectation of reward decreased, due to the imprecise timing mechanism. Specifically, for sufficiently long pre-reward periods the RPEs exhibited a tendency to increase toward the end of the possible reward window. The exact profile of the RPEs across different ISIs depended on the precision of the temporal representation. The results of the simulations showed that a SH representation of average precision ($k = 40$) could cope with all the current data. Particularly noteworthy was the prediction of a more pronounced RPEs for long ISIs when the temporal parameters of the simulation were similar to the experimental condition reported in (Bromberg-Martin et al., 2010).

The reset mechanism has been introduced in the TD machinery by (Suri and Schultz, 1998, 1999; Brown et al., 1999) to fix some mispredictions of the original model formulation proposed in (Montague et al., 1996). In particular this original formulation produced a spurious inhibition at the usual time of reward occurrence for rewards delivered early than expected, a prediction inconsistent with the data reported in (Hollerman and Schultz, 1998). The reset mechanism has been criticized for being an ad hoc device of doubtful generality, and for being not revealing about the computations carried out by dopaminergic neurons (Daw et al., 2006). It has been instead suggested that dopaminergic RPEs were shaped by state uncertainty. In this perspective time is nothing more than any other state that cannot be observed directly (i.e a hidden state), and that the system needs to infer in order to calculate reward predictions and reward prediction errors. Such a hidden state approach is compelling because it proposes a normative solution to the problem of keeping track of time in RL models, that can be achieved by performing TD learning on the state posteriors (the belief state; see subsection 1.1.5). Additionally, it has been further supported by a recent study in which reward delivery occurred only probabilistically (Starkweather et al., 2017). However it is interesting to note that the hidden state approach provides, in limited circumstances, a justification to the reset mechanism. In this scheme the occurrence of the reward (or of a reward-related event) induces indeed a state transition and therefore the reset mechanism arises as a consequence of the inference process.

In this thesis I will not tackle the problem of inference about time . This approach is dictated by the fact that the central part of the thesis will focus on a model based analysis of the dopamine activity recorded while monkeys were engaged in complex decision making task. In the context of decision making animals face the non-trivial problem of

extracting information about ambiguous stimuli in order to perform rewarding choices. In the formal theory this means that the the agent-environment interaction can be modelled as a POMDP, in which the properties of the stimuli relevant to achieve future rewards correspond to the state that needs to be inferred. I will therefore concentrate on this second form of inference (because of its relevance in DM), keeping aside the problem of inference about time. Nevertheless, a timing mechanism is unavoidable in RL models of the dopamine system. Thus, guided from the preliminary analysis of this chapter, I will thus assume in what follows a SH representation of average precision and the presence of the reset mechanism. Integrating state and time inference in a formal theory of RL and DM of the dopamine system is an intriguing problem that could be the object of future research, but that remains beyond the scope of this thesis.

Chapter 3

The dopamine signal in tasks with sensory and temporal uncertainty

3.1 Introduction

When an inexperienced animal hears a soft rustle in the nearby foliage it does not associate this cue with the escaping prey that it observes immediately after. How does the animal get to learn that that sound is a predictor of a possible catch and that the correct action to take is to approach it and try to get it? In perceptual decision-making experiments, animals learn how to make decisions based on their perception of weak sensory stimuli, receiving a reward for their correct choices, which they are taught to communicate by means of a specific motor action (Hanes and Schall, 1996; Shadlen and Newsome, 1996; Romo et al., 1999; Shadlen and Newsome, 2001; Cook and Maunsell, 2002; de Lafuente and Romo, 2005; Hernández et al., 2010). The learning of these tasks is presumably mediated by the activity of midbrain dopamine (DA) neurons (Schultz, 1998). Although DA recordings made while animals are engaged in making such difficult decisions are scarce, experiments on Pavlovian and instrumental conditioning have shown that under a novel stimulus-reward association, DA neurons respond to the unexpected reward with an activity burst. Remarkably, after training this phasic response is shifted to the conditioned stimulus where it works as a signal predicting the future reward (Romo and Schultz, 1990; Ljungberg et al., 1992; Mirenowicz and Schultz, 1994; Hollerman and Schultz, 1998; Schultz, 1998). From a computational standpoint, reinforcement learning (RL) methods (Sutton and Barto, 1998) have been successfully applied to explain this and many other observations (Schultz et al., 1997; and for reviews see: Niv, 2009; Maia, 2009; Ludvig et al., 2011). According to the reward prediction error hypothesis (RPE) (Barto, 1995;

Montague et al., 1996), the DA phasic activity signals an error in the prediction of the expected total reward (Bayer and Glimcher, 2005; Steinberg et al., 2013; Chang et al., 2016) and it is used to learn associations between rewards and task events.

In classical and instrumental conditioning the reward acts as a reinforcer strengthening the association with the stimulus, provided the animal follows the task instructions. In some experiments the reward was delivered only after the animal made a choice between alternative options(Morris et al., 2006; Roesch et al., 2007; Bayer and Glimcher, 2005). However, in those studies the task events were unambiguous: the perceptual reports were always correct and there was a well-defined temporal relationship between the perceived stimulus and reward delivery. But, this is very different from the real world situation described above in which the reward is announced by a muted sound produced in a noisy environment at an unexpected time. Consequently, little is known about the DA signal in such uncertain conditions and up to now few experiments have attempted to fill this gap. The existing studies seem to indicate that the DA signal has a much richer structure than in simple choice paradigms. For example, Nomoto et al., 2010 found that the response to visual dynamic random dot stimuli is more complex than the response to the stimuli commonly used in previous studies. The DA activity seemed to follow a more elaborate temporal profile, first responding abruptly to the onset of the stimulus (presumably due to its detection) and then producing a more extended response (supposedly due to the decision-making process, Nomoto et al., 2010; see also Schultz, 2015). In another recent study, de de Lafuente and Romo, 2011 recorded DA neurons while a monkey was engaged in the detection of weak vibrotactile stimuli. In this task, when the animal was instructed to communicate its choice by pushing one of two push buttons, these neurons coded the uncertainty associated with a perceptual judgement about the presence or absence of the stimulus.

Here, we combined data analysis and computational modeling to investigate the DA signal recorded as monkeys detected weak vibrotactile stimuli applied at random times (Figure 3.1a and Methods). In this task (de Lafuente and Romo, 2005, 2006), a start cue indicated the beginning of a trial and was followed by an interval of variable duration after which, with probability 0.5, the vibrotactile stimulus was applied. After a fixed interval, a go cue instructed the monkey to communicate its decision about the presence or absence of the stimulus by pushing one of two buttons. The animal was rewarded in all correct trials. The difficulty of the task stems from the use of very weak stimulus amplitudes and from the uncertainty about the time of possible stimulation. It has been proven that because of these uncertainties, the firing activity of frontal lobe cortex neurons

codes internal processes associated with the elaboration of the decision reports in this task (Carnevale et al., 2012, 2013, 2015). A key result in the midbrain DA system was that the neurons' response to the go cue is weaker in trials with stimulus-present choices (hit and false alarm trials) than in trials with stimulus-absent choices (correct rejection and miss trials) (de Lafuente and Romo, 2011). This was attributed to the higher certainty that the animal has on its decision in "yes" response trials. The result is important because it indicates that the DA phasic response reflects internal processes; however, several issues have been left unanswered. For instance, the nature of those processes was attributed to decision certainty on the basis of a comparison of the probabilities of reward in stimulus-present versus stimulus-absent trials, which resulted higher in the former case. However, in the task the animal made a choice and received a reward only after the delivery of the go cue. It is then not clear whether the DA phasic response to that signal was related to the choice itself or to some other process that occurred during the formation of the decision. Besides, whatever the nature of the process, it should be explained why it became visible in the DA activity under the application of the go cue. Finally, the response to this event was different in each of the four trial types and the reason of this graduation in the DA activity is not known.

Apart from stimulus uncertainty the detection task also has temporal uncertainty. The effect of the trial-to-trial variability in the trial duration on the DA activity was not considered in the previous work. However, it is known to have important consequences over prefrontal neurons (Carnevale et al., 2012, 2015) and it is reasonable to believe that it will also affect the midbrain DA system. In fact, effects of temporal variability on DA neurons have been reported several times in tasks without stimulus uncertainty (Fiorillo et al., 2008; Bromberg-Martin et al., 2010; Pasquereau and Turner, 2015) or with it (Nomoto et al., 2010). To investigate these issues we have taken a different approach, proposing a model based on the RL framework and using it to understand the activity of DA neurons. Because of the two kinds of uncertainty, on the stimulus amplitude and trial duration, the model estimates the total reward and reward prediction errors using belief states (Dayan and Daw, 2008; Rao, 2010; Bogacz and Larsen, 2011).

3.2 Results

3.2.1 Temporal profile of the DA response

Behavior can be described in terms of the four possible trial types of the vibrotactile detection task. Stimulus-present trials can be correct (hits) or wrong (misses) responses, while stimulus-absent trials produce correct rejection (CR) or false alarm (FA) responses.

Reward is delivered only in trials with correct responses. Since we want to discuss the data in the framework of RL we kept only those neurons compatible with RL ideas; all the electrophysiological results presented in this work have been obtained from midbrain DA neurons responding to reward delivery with a positive phasic activation in correct (rewarded) trials and with a pause in error (unrewarded) trials. These are 23 out of the 69 neurons analyzed in de Lafuente and Romo, 2011 (see Figure B.1 and Methods about the selection criteria). Figure 3.1b shows the average firing rate of the population of the selected DA neurons during the vibrotactile detection task. This temporal profile is similar to that of the firing rate of the larger population of midbrain DA neurons analyzed before (de Lafuente and Romo, 2011). However, there seems to be an important difference between the two datasets: in Figure 3.1b, the DA activity immediately before the go cue exhibits a pronounced decay in all trial types. Instead, the firing rate of the discarded neurons does not show this modulation (Figure B.2).

3.2.2 Transient DA activity in the period of possible stimulation

If the DA response to the go cue codes some type of certainty we wondered how the activity of the DA midbrain neurons might have acquired this property. We reasoned that a detection process and the certainty about detected events could have been elaborated in cortical circuits and then transmitted to midbrain neurons producing transient changes in their activity. We have then investigated the existence of transient activation of the DA neurons during the possible stimulation window (Figure 3.1a; the interval between 1,5 s and 3,5 s after the key down event).

It has been suggested that the initial response of DA neurons to external stimuli reflects their physical salience (Schultz, 2015). In fact, Figure 3.1b shows that in hit trials the vibrotactile stimulus generates a clear transient response with a linear dependence of the neurons' firing rate at the stimulus onset as function of the stimulus amplitude Figure 3.1c. This effect had been observed for the larger dataset (de Lafuente and Romo, 2011) but here we show that it is also present for neurons compatible with the RL framework.

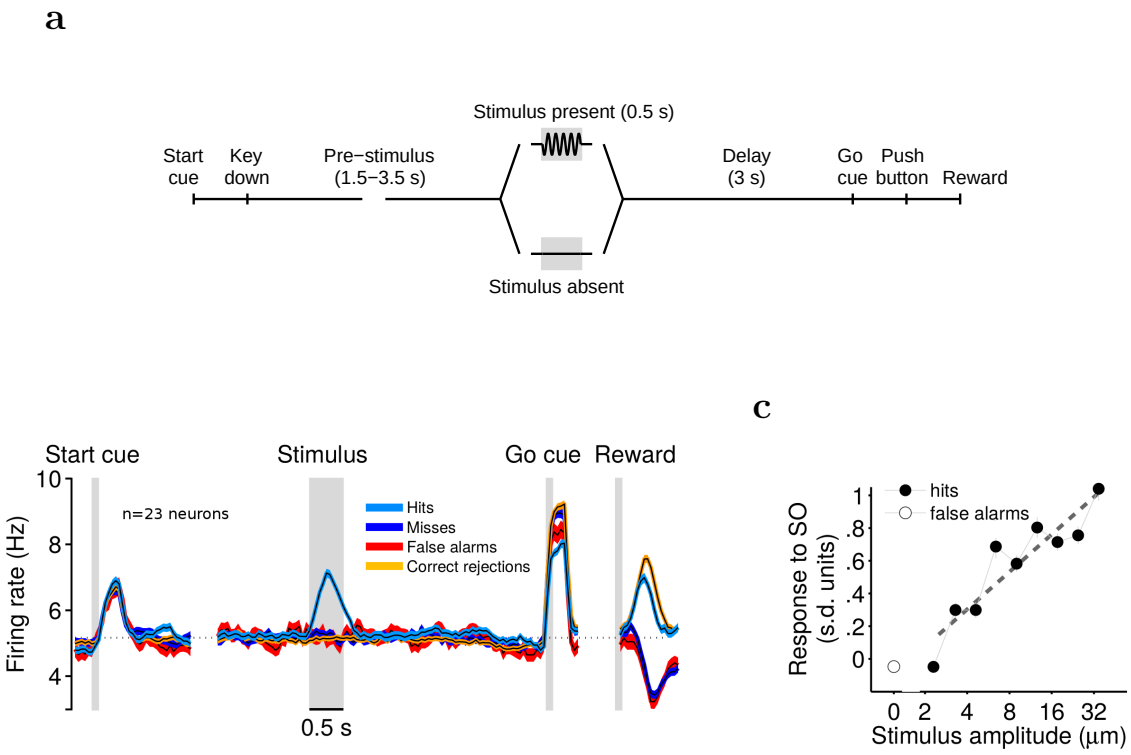


Figure 3.1: **Detection task and temporal profile of the DA neurons activity.** (**a**) Trials began with a start cue instruction, i.e when the stimulator probe indented the skin of one fingertip of the restrained hand. The monkey reacted by placing its free hand on an immovable key (key down event). In stimulus- present trials, after a variable pre-stimulus period (1.5-3.5 s), a vibratory 0.5 s stimulus was presented. Then, after a fixed delay period (3 s), the go cue (stimulator probe tip lifted off from the skin) was delivered and the monkey communicated its decision by pressing one of the two buttons (push button event). The reward was delivered immediately after the push button event in correct choice trials. Stimulus-absent trials had the same temporal structure with the only difference that the vibrotactile stimulus was not presented. (**b**) Mean population firing rate (black line, ± SEM colored bands) plotted as a function of time for the four trial types. Activity is aligned to the start cue (left), go cue (center) and reward delivery (right). The dotted line indicates the baseline activity (5.1 spikes per second). Before the go cue the activity exhibited a pronounced decay with respect to the baseline in all trial types. (**c**) Responses of DA neurons at the stimulus onset (SO) in yes-decision trials sorted by stimulus amplitude. Data showed a positive linear increase of the response with the amplitude of the stimulus ($R^2 = 0.98$, $P < 0.001$) (See Methods for more details on data analysis).

Figure 3.1b suggests that the vibrotactile stimulus generates a phasic response in the DA neurons only in hit trials (de Lafuente and Romo, 2011). However there are reasons to believe that the apparent unresponsiveness of DA neurons in FA and miss trials requires a more detailed analysis. For example, in FA trials the animal indicated the presence of a stimulus and this perception could somehow be reflected in the activity of DA neurons. Also, since the majority of miss trials occur for low amplitude stimuli, the existence of a transient response to high amplitude stimuli might be hidden in the mean over all miss trials (neurons in cortical areas are activated by the stimulus even in miss trials, de Lafuente and Romo, 2005). Hence, one should not discard that in high amplitude miss trials the information about the presence of a stimulus is transmitted to midbrain neurons. We have then investigated whether there are transient DA responses in high amplitude miss trials and FA trials.

In miss trials the onset of the stimulus seems not to produce any evident modulation of the firing rate (Figure 3.1b); it could then be argued that in these trials the stimulus was not detected by cortical frontal neurons. Indeed, most miss trials occur when the stimulus amplitude is weak and the firing rate of DA neurons is not modulated by its application (green trace in the left panel of Figure 3.2a). However, when high amplitude miss trials are analyzed, we see that the firing rate of the cells did increase at stimulus onset (blue trace in the left panel of Figure 3.2a).

In FA trials, although the subject reported the presence of a stimulus, the firing rate in Figure 3.1b does not show any apparent modulation. Thus, it is not clear how a stimulus-present choice was elaborated during the trial. A recent work about frontal lobe cortex neurons recorded while monkeys performed the same detection task (Carnevale et al., 2015) sheds light on this issue. In FA trials, those cortical neurons underwent transient activity increases resembling the response to a weak true stimulus. These transient FA events occurred at random times within the possible stimulation window; that is, inside the 2 s-interval starting 1.5 s immediately after the key down event (Figure 3.1a). We have then assumed that these events are transmitted to DA neurons in a way similar as true stimuli are. If this assumption were correct, then the mean firing rate of DA neurons in FA trials, computed during the possible stimulation period, should be slightly higher than the mean firing rate in CR trials evaluated during the same period. To test this hypothesis, we aligned all FA trials to the key down event and compared their mean firing rate in the possible stimulation window with the mean firing rate of CR trials computed in the same temporal window. The results indicate that the mean firing rate in FA trials is significantly higher than in CR trials (left panel of Figure 3.2b). This seems to be an

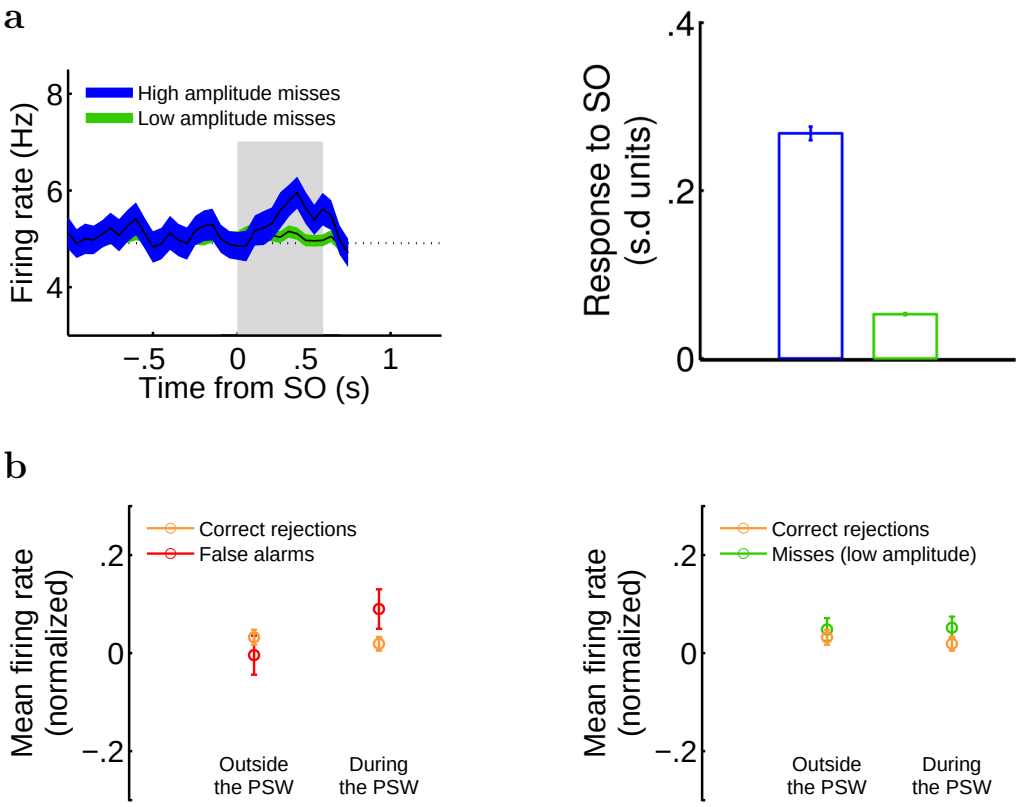


Figure 3.2: **Signatures of detection in FA trials and high amplitude miss trials.** (**a**) In high amplitude miss trials DA neurons responded transiently to the vibrotactile stimulus (left). The activity of neurons after the stimulus onset (SO), standardized with respect to a pre-stimulus window (see Methods), showed a significant phasic activation ($P < 0.05$, two sample t test) in high amplitude miss trials compared to low amplitude ones (right). (**b**) The mean activity in FA trials (see Methods) exhibited a significant positive modulation with respect to CR trials during the possible stimulation window (PSW) ($P < 0.05$, two sample one-tailed t test) but not outside it ($P = 0.80$, two sample one-tailed t test) (left). On the contrary the activity in low amplitude miss trials was indistinguishable from correct rejection ones both outside ($P = 0.28$) and within ($P = 0.11$) the PSW (right).

exclusive property of this particular temporal interval: the mean firing rates in FA and CR trials computed outside the possible stimulation window are rather similar (left panel of Figure 3.2b). As a further test that the elevation of the firing rate during the possible stimulation period is specific to FA trials, we did a similar analysis with low-amplitude miss trials aligned to the key down event (Figure 3.1a). In contrast to what happened with

FA trials, the mean firing rate in low amplitude miss trials was not significantly different from that of CR trials neither within the possible stimulation window nor outside the possible stimulation window (right panel of Figure 3.2a).

The transient events discussed above are presumably related to **detection processes** taking place prior to their reception by midbrain neuron and much before the animal reports its choice. We interpret them as contributing to the **certainty about the detection** of transient activity fluctuations in circuits presynaptic to the midbrain DA system, distinguishing it from **certainty about the choice**, a term which should be used after the animal indicates its decision (Pouget et al., 2016). A precise definition of certainty about the presence of the stimulus will be given later, in the context of our RL model; however we can give a qualitative argument explaining why the transient events contribute to this certainty. Regardless of whether the transient activation was produced by a true stimulus (as in hit and high amplitude miss trials) or by some internal process (as in FA trials), the transient event works as a subjective confirmation that a stimulus was detected and hence it increases the certainty about its presence. The degree to which the transient event contributes to the certainty would depend on its strength. For instance, transient events generated in FA trials have a similar effect on DA neurons as those produced by true low amplitude stimuli (see Figure B.4) and they could convey a similar level of certainty about their detection.

3.2.3 Salience of the go cue and effects of temporal uncertainty

To obtain further insight about how the response to the go cue acquired a dependence on the certainty about the presence of a vibrotactile stimulus we investigated the effect of the stimulus amplitude on this task event. First, we notice the DA response to the go cue decreases linearly with the amplitude of the stimulus (left panel in Figure 3.3); this is similar to the results found previously for the larger dataset (de Lafuente and Romo, 2011). For the moment we do not make any interpretation about this result, preferring to discuss it in the context of the model presented below. Instead, we now wonder whether the response to the reward delivery also exhibits a dependence on the stimulus amplitude. The analysis shows that the dependence disappears (right panel in Figure 3.3). Our interpretation of this observation is that the go cue acts as a physically salient signal erasing from the DA activation (at least partially) the dependence on the properties of previous task events. In fact, the responsiveness of DA neurons to the physical salience of stimuli has been discussed frequently (see, e.g., Schultz, 2015).

The effects of the trial-to-trial variability in the duration of the interval immediately

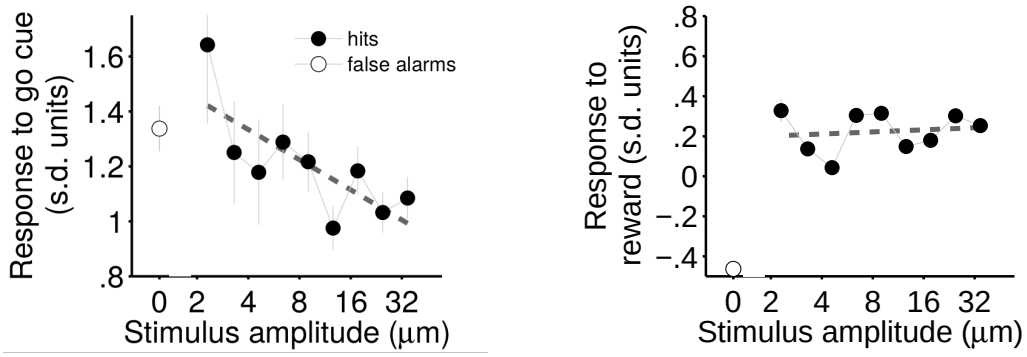


Figure 3.3: **Responses to the go cue and to the reward delivery.** In stimulus present decision the DA response at the go signal linearly decreased with the amplitude of the stimulus (left). The dependence on the amplitude completely disappeared in DA activity at the reward delivery ($R^2 = 0.58$, $P = 0.02$ for the linear regression at the go cue; $R^2 = 0.02$, $P = 0.72$ for the linear regression at the reward delivery). For details about the standardized responses to the go cue and to the reward delivery see Methods.

before the go cue are visible in the phasic DA responses to that event. The data analysis shows that, in both CR (Figure 3.4a, left) and low amplitude miss trials (Figure 3.4a, right), the longest the key down-go cue interval, the stronger is the DA phasic activation. This is opposite to what was found in some other studies (Fiorillo et al., 2008; Nomoto et al., 2010) but agrees with (Bromberg-Martin et al., 2010). We will come back to this issue later, when we will explain this result with our RL model. In contrast, the response to the delivery of reward is the same for long-duration and short-duration trials, both in CR and low amplitude miss trials (Figure 3.4b).

The variability of the duration of the trials also produces a modulation of the DA activity during the period previous to the go cue (Figure 3.1b). To analyze this effect, we have aligned CR trials at the key down event. The resulting firing rate has a negative modulation starting at about the predicted time (Figure 3.4c, top). We then asked whether low amplitude miss trials, when aligned to the key down event, showed a temporal profile similar to CR trials. Indeed, in low amplitude miss trials, the DA phasic response to the stimulus was not present (Figure 3.2a), and the mean firing rates inside and outside the possible stimulation window are not significantly different (right panel of Figure 3.2b).

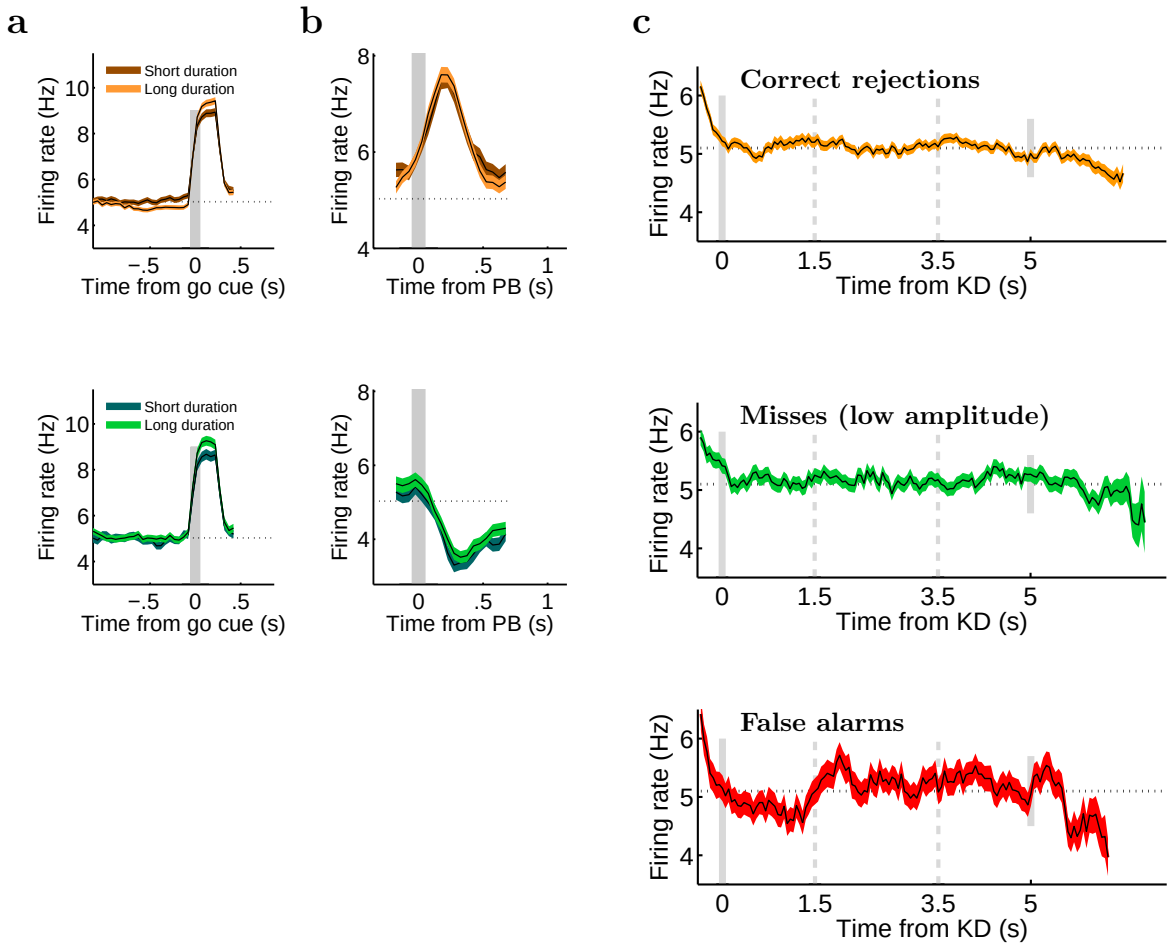


Figure 3.4: **Temporal Expectation.** (**a**) Trial duration modulated the phasic DA response to the go cue but did not affect its response after the reward delivery. Left: When CR trials were sorted according to the duration of the key down event-go cue interval the DA response to the go signal resulted stronger for long duration trials (right). Right: the same effect is observed for miss trials of low amplitude. The gray lines indicate the go cue. (**b**) The response of neurons to the reward delivery was independent of the trial duration. Left: CR trials. Right: miss trials of low amplitude. The gray lines indicate the push button event (PB). (**c**) In the four trial types before the go cue the DA firing rate showed a slow negative modulation (Figure 3.3a). Top: The mean population activity of DA neurons in CR trials aligned to the key down event (KD). The activity started to decrease around the first time when the go cue could appear (gray short line on the right). This negative deviation from the baseline increases as the time elapses and the go instruction becomes more and more expected. Middle: the same effect is observed in low amplitude miss trials. Given the lack of response to the stimulus presentation in this fraction of miss trials, the DA activity anticipated the go cue presentation similarly to what was observed in CR trials. Bottom: FA trials exhibited a similar temporal behavior.

Clearly, their alignments to the key down event are also comparable; starting about 2 s immediately before the go cue, the dopamine signal exhibits a tonic negative modulation, as the one quantified in CR trials (Figure 3.4c, middle). The same effect is seen for FA trials (Figure 3.4c, bottom).

3.2.4 The reinforcement learning model: formulation

It has been suggested that when the brain does not have full access to the correct value of the physical attributes of the stimuli, the cerebral cortex uses noisy observations to infer them (Knill and Richards, 1996; Rao et al., 2002), and that the midbrain DA neurons and striatal neuronal circuits evaluate the state of the environment to select the appropriate actions based on the results of that inference (Dayan and Daw, 2008; Rao, 2010; Bogacz and Larsen, 2011). In this scheme, the outcome of the inference process is a posterior probability about the state of the environment, which is interpreted as a measure of the belief about that state (Kaelbling et al., 1998). In line with these ideas, we have assumed that a Bayesian module (representing cortical circuits) accumulates sensory evidence to compute a time-dependent posterior probability about the presence of the vibrotactile stimulus (hereafter referred to simply as the belief and denoted as $b_{sp}(t)$). The belief is then sent to a RL module, representing midbrain DA neurons and striatal neuronal circuits (Figure 3.5a). This is a valuation and action selection module that makes predictions about the future reward, computes the error of this prediction (the RPE) and chooses whether to press one button indicating a stimulus-present choice or another button press indicating the stimulus-absence choice.

A crucial question is: When and how does the outcome of the accumulation process (the belief state) affect the reward prediction and action selection operations? In the analysis of the experimental data we found that DA neurons are activated transiently in hit trials and in high amplitude miss trials by the vibrotactile stimulus and also in FA trials during the window of possible stimulation by a stimulus-independent process. A simple and plausible assumption is that the events responsible for those activations are related to a belief evaluated by cortical circuits that exceeded a threshold value (the maximum a posteriori -MAP- criterion sets the threshold at 0.5). Specifically, in the model we assume that when the belief computed by the Bayesian module grows beyond that threshold it is sent to the relevant downstream structures. When this happens a representation of the stimulus is turned on in the RL module and it is used to establish associations between the reward and the stimulus. To accomplish this function, the RL module operates following RL rules based on belief states with two other important additions, inspired from the

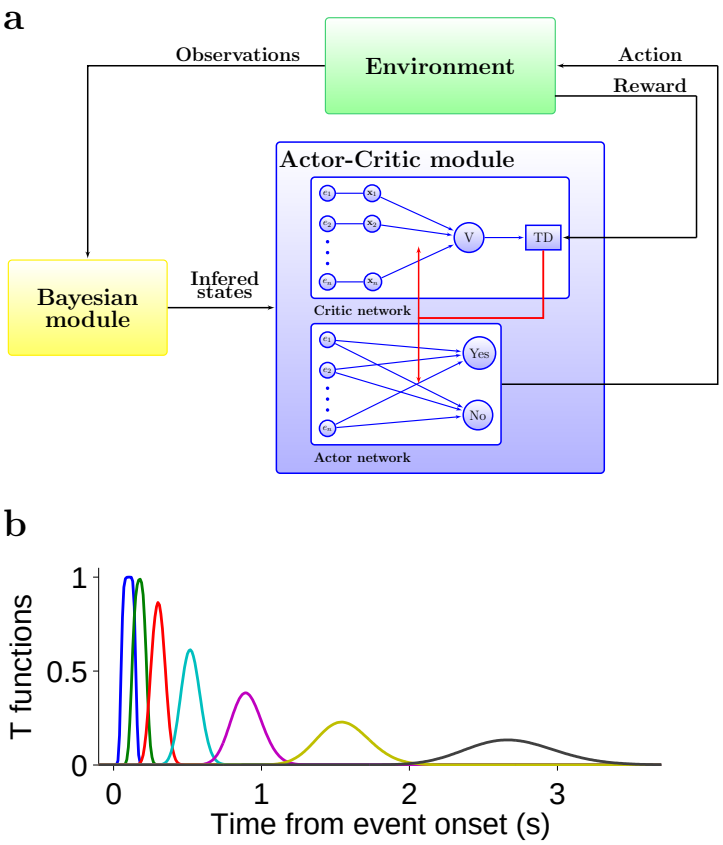


Figure 3.5: **Model architecture and temporal representation of task events..** (**a**) The model relied on two structures: a Bayesian module and a reinforcement learning module. The Bayesian module used the noisy observations received from the environment to compute a time-dependent posterior probability (the belief) about the presence of external events and sent it to a RL module. The RL module consisted in an actor-critic architecture (Sutton and Barto, 1998). It used the information inferred by the Bayesian module to evaluate and to select actions (see Methods for more details on the model). (**b**) Each task event was represented across time via a set of functions reproducing the event at different latencies from its onset. Importantly, the resolution of the representation degraded with the passage of time (Shankar and Howard, 2012).

previous data analysis. First, on the basis of the physical salience of the go cue observed in the data (Figure 3.3), we introduce in the RL module a reset mechanism that allows events predicting a high reward to disrupt the internal representations of earlier events (Suri and Schultz, 1999). This mechanism does not introduce any parameter in the model; see Methods).

Second, given the effects of the variable duration of the trials found with the data analysis (Figure 3.1b and Figure 3.4), each task event is represented with a temporal resolution that degrades with the passage of time (Shankar and Howard, 2012) (Figure 3.5b). To update the value of states and actions the RL module computes the error made in the prediction of the reward as $\delta(t) = r(t) + TD(t)$, where $r(t)$ is the reward received at time t in a trial and TD(t) is the temporal difference between the total rewards predicted at times $t+1$ and t (Sutton and Barto, 1998; see Methods for further details on the model). According to the RPE hypothesis, $\delta(t)$ should be compared with the population average of the mean firing rate of the DA neurons.

In the following we will use the model to show that these mechanisms, belief states, transmission of detected events, salience and a temporal representation with limited resolution, suffice to explain the DA response to the go signal. We start by verifying that the salience of the go signal actually produces a RPE after the reward delivery independent of the stimulus amplitude. Then we analyze the events transmitted from the Bayesian to the RL module in hit trials, high amplitude miss trials and FA trials. After that, we present the model explanation of how belief states produce a RPE at the go cue that depends on the trial type, reproducing the graded response of the DA neurons to this task event. The explanation of this observation is one of the main objectives of the proposed model. Finally, we close the analysis of the model with a study of how temporal expectation modulates the RPE and propose an explanation of the differences in various experimental observations about the dependence on trial duration of phasic responses.

3.2.5 Reset of the go cue

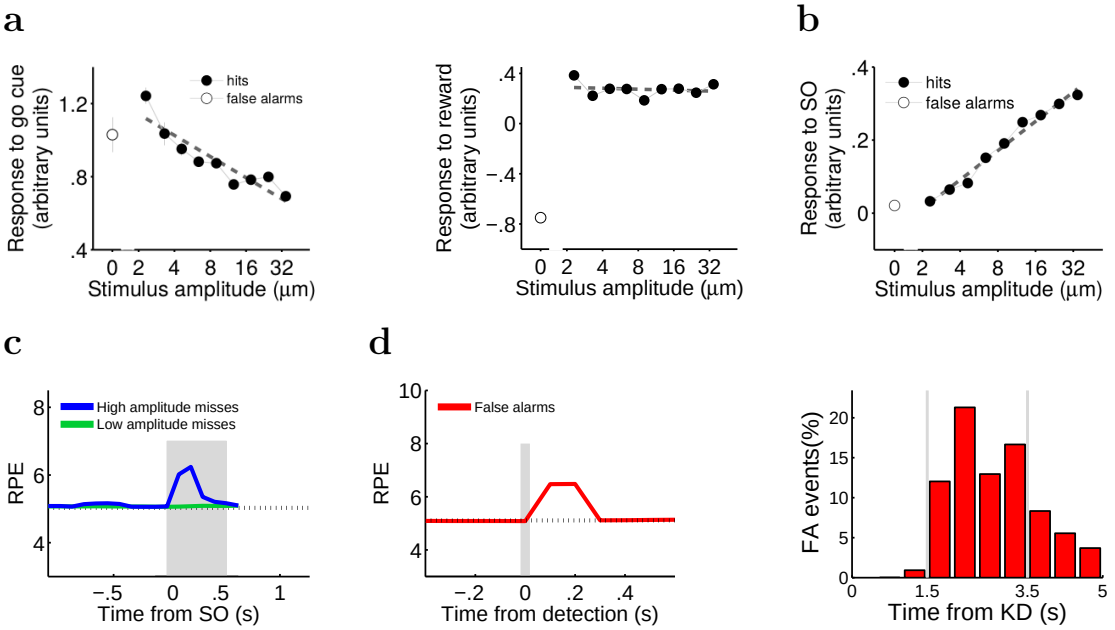
Data show that although the DA phasic activation at the go cue depends on the stimulus amplitude this dependence disappears in the response to the reward (Figure 3.3), supposedly as a consequence of the physical salience of the cue signal. This property led us to formulate a RL model in which the task events are endowed with a reset mechanism. We now analyze in the model the effect of this mechanism on the dependence on the stimulus amplitude of the RPE at the go cue. In agreement with the data (Figure 3.3), numerical simulations of the model exhibit a decreasing linear dependence of the RPE at the go cue with the stimulus amplitude (Figure 3.6a, left) whereas after the delivery of the reward the analysis does not show a significant slope (Figure 3.6a, right).

3.2.6 Transmitted events during the period of possible stimulation

We start by verifying that the belief transmitted from the Bayesian to the RL module produces transient changes in the RPE in correspondence to those observed in the DA activity. In hit trials, after the application of the vibrotactile stimulus, the RPE increases linearly with the stimulus amplitude (Figure 3.6b), as the DA response does (Figure 3.1c).

An immediate prediction of the model is that miss trials can arise in two possible ways. The most frequent ones happen when the stimulus is too weak to be detected by the Bayesian module. Less often, for stronger amplitudes, even if the stimulus is detected (Figure 3.6c) the variability of the action selection process may generate a stimulus-absent choice, an effect that in our simulation occurred in about 12% of all miss trials. In fact, data show that in high amplitude miss trials the animal reported stimulus-absence, though the firing rate of the cells did increase at stimulus onset (blue trace in the left panel of Figure 3.2a). Similar mismatches between the cortical detection and action selection outcomes are also present in CR and FA trials (table in Figure 3.6e).

In the RL model, the times when the belief exceeds its threshold value are known. FA trials aligned to those times evidences a transient increase of the RPE signal $\delta(t)$ at the time of the FA events (left panel of Figure 3.6d). Furthermore, those detection times are distributed mainly during the possible stimulation window (right panel of Figure 3.6d). Interestingly, the distribution is similar to the one found from the activity of prefrontal neurons (Carnevale et al., 2015). Notice that in FA trials perception arises from detected transient events in the cortical module followed by a "yes" response.



e

	Stimulus detected by the bayesian module
Hits	≃ 100 %
Misses	≃ 12 %
CRs	≃ 1 %
FAs	≃ 84 %

Figure 3.6: **Basic properties of the RPE.** (**a**) The RPE at the go cue depended on the stimulus amplitude but this dependence was lost at the reward delivery. In stimulus-present decisions the RPE at the go cue linearly decreased ($R^2 = 0.84$, $P < 0.001$) with the amplitude of the stimulus (left). The dependence on the amplitude completely disappeared in the RPE at the reward delivery ($R^2 = 0.03$, $P = 0.64$) as a consequence of the reset property of the go signal (right). This has to be compared with the DA activity in Figure 3.3. (**b**) Responses at the stimulus onset (SO) in yes-decision trials as predicted by the model sorted by stimulus amplitude. The model showed a positive linear increase of the response with the amplitude of the stimulus ($R^2 = 0.98$, $P < 0.001$). See Methods for more details on the model analysis). (**c**) The model predicted a response to the stimulus as a consequence of a Bayesian detection in miss trials when the amplitude is high. A similar response was apparent in the data Figure 3.2b. (**d**) The RPE in FA trials after an erroneous detection showed a phasic response (left). In the model these erroneous detection events were produced mainly within the PSW (right). KD denotes the key down event. (**e**) Percentage of trials where a transient event was detected by the Bayesian module, for each of the four task contingencies. Note how the occurrence of a detected event in the Bayesian module did not by itself generate perception (miss trials). The values of the model parameters are given in **??** in Methods.

3.2.7 Certainty about the presence of the stimulus

We now turn to the main issue we want to address with the model: how a RL module receiving uncertain information through a Bayesian inferring process can explain the graded phasic response to the go cue. The computations carried out by the DA neurons during the delay period are crucial to understand and interpret their responses to the go cue. The immediate effect of a large stimulus belief on the RL module is to initiate the evaluation of how much reward it predicts till the end of the trial; that is, the estimated value of the stimulus. Figure 3.7a shows the reward predicted by the stimulus in trials with stimulus-present choices. The predicted reward increases in a graded manner with the stimulus amplitude. The graduation is maintained during all the delay period, until the presentation of the go cue. Note that the transient events in FA trials predict a reward similar to that estimated by low amplitude stimuli (red line in Figure 3.7a). The predicted reward increases with time during the delay period, as a consequence of the temporal discount.

The total predicted reward in trials with stimulus-absent choices lie below those estimated in trials with stimulus-present choices (Figure 3.7b). This is because in the first case the key down event is the only task event contributing to the prediction, in the second case the detection of an event increases the belief about the presence of the stimulus so that it can reach its threshold and generate an extra contribution. Another relevant observation is that the predicted reward is slightly higher in miss than in CR trials. As we noticed before, miss trials behave as CR trials, but only when low amplitude stimuli are presented; for high amplitude stimuli, a detected event increases the belief which then is transmitted from the cortical to the RL module producing a somewhat higher estimated value.

When the go cue is applied its high physical salience partly erases the information about the stimulus amplitude and the corresponding reward predictions collapse in approximately the same value (Figure 3.7a). Since the error of each of these predictions at the time of the go cue, $\delta(t)$, is the difference between the reward predicted by this event and the prediction at the time preceding it, the response to the go cue should be higher in FA than in hit trials, a result which is verified by the data (compare the RPE in Figure 3.7c with the response of DA neurons to the go cue in Figure 3.1b). A similar argument explains why the response to the go cue in CR trials is slightly higher than in miss trials (Figure 3.1b and Figure 3.7c); here the small difference comes from the higher value of miss trials during the delay period (Figure 3.7b). Finally, since during the delay period the predicted reward in trials with stimulus-absent choices is smaller than in trials

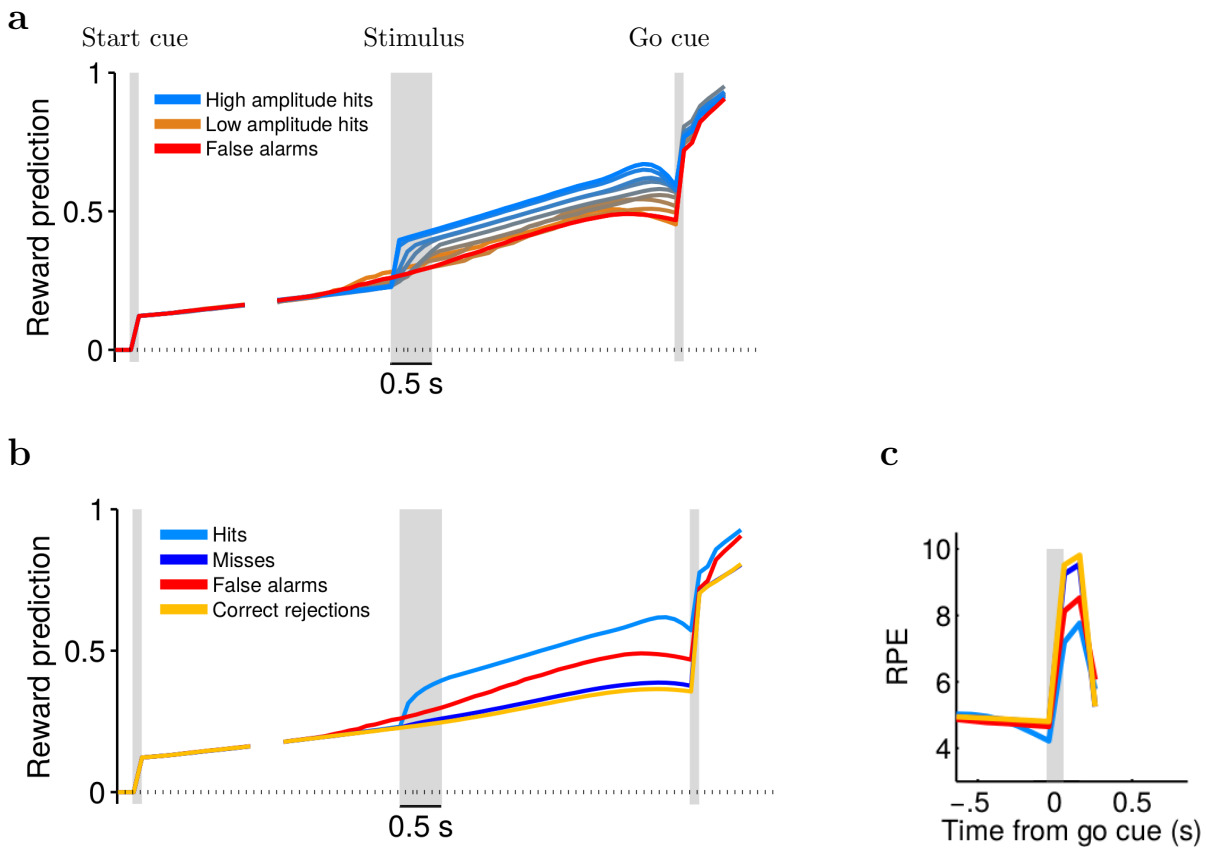


Figure 3.7: **Predicted reward during the delay period and responses to the go cue and to the reward delivery.** (**a**) In trials with stimulus-present choices during the delay period the predicted reward increased with the stimulus amplitude in a graded manner. In FA trials (red line) its temporal profile was similar to that observed when a low amplitude stimulus is perceived. (**b**) The predicted reward during the delay period resulted higher in trials where the bayesian module detected a stimulus. It was greater in miss than in CR trials due to the detection of the stimulus when the amplitude was high. At the go cue, because of the reset mechanism, the reward predictions in the four trial types collapsed in approximately the same value and immediately after they separated in two values corresponding to the possible decisions. (**c**) The RPEs at the go cue were lower on stimulus-present decisions (hit and FA trials) than in stimulus-absent choices (miss and CR trials). According to the model this graduation was determined by the modulation of the reward prediction as described in **a** and **b** and by the reset mechanism.

with stimulus-present choices, the responses to the go cue are larger in the former case than in the latter. The resulting model prediction for the response to the go cue in the four trial types is summarized in Figure 3.7c.

The above arguments explaining the response to the go cue can be phrased in terms

of how the subject's certainty about a detected event evolves throughout the delay period. This certainty can be defined as the probability of a correct detection. Since the Bayesian module decides about the presence of a stimulus using the MAP criterion, the probability of a correct cortical detection is either the posterior probability about the stimulus-present state (i.e. the belief $b_{sp}(t)$), if this posterior is above 0.5, or the posterior about the stimulus-absent state (i.e. $1 - b_{sp}(t)$), if it is below 0.5. When the Bayesian module transmits the belief to the RL module, we can then say that all the subsequent computations done in this module are based on the certainty that the received information is correct. In particular, the different responses to the go cue in hit and FA trials are due to the difference in certainty of these two trial types. Also, the difference between the responses in miss and CR trials comes from the higher level of certainty in a fraction of miss trials. The smaller response to the go cue in stimulus-present choices than in stimulus-absent choices can be attributed to the larger certainty of the animal when it reports the stimulus presence.

The go cue predicts the total future reward averaged over "yes" and "no" responses. After its delivery, because of the reset mechanism, the RPE ceases to depend on the trial type and starts coding the possible choices. This is seen in the response to the reward both in the model and the data (Figure B.3). The smaller RPE in hit trials than in CR ones is explained by a larger fraction of rewarded trials of the former type.

3.2.8 Temporal expectation

In the detection task used here, the time sequence of some events is not fixed and their presentation cannot be predicted. Studies in cortical areas indicate that this produces an expectation of the forthcoming events that is governed by the subjective hazard of occurrence of the expected event (Janssen and Shadlen, 2005). This temporal expectation might affect DA neurons by modulating their firing rate during the intervals between task events (Fiorillo et al., 2008; Nomoto et al., 2010; Bromberg-Martin et al., 2010). In our detection task this is particularly evident during the interval preceding the go cue, where the firing rate in the four contingencies decreases with respect to its baseline value (Figure 3.1b). Note, however, that the duration of this interval depends on the trial type. While in stimulus-present trials the delay period has a fixed duration (3 s), in stimulus-absent trials the interval between the key down event and the go cue varies from trial to trial taking values between 5 and 7 s (Figure 3.1a). However, the fact that the decay is also observed in hit trials with a fixed stimulus onset-go cue interval suggests that there must be other factors responsible for the decrease of the firing rate. According

to the model, the possible causes are: in some hit trials, particularly those with weak stimulus amplitude, the event detected by the Bayesian module was not the stimulus itself but a noisy fluctuation, similar to what happens in FA trials. In these trials, the effective duration of the delay period depends on the time when the fluctuation occurs, which lies within a 2 s temporal window (Carnevale et al., 2015). However, these are only a small fraction of the total number of hit trials and this effect is expected to give a small contribution. Even rarer are those weak amplitude trials in which the stimulus was not detected, but variability in the selection of the action led to the correct response. Finally, an imprecise estimate of the duration of the delay period could also lead to an effective variability of this interval. This effect occurs in all trials and it could be the most important explanation of the decaying tonic activity in hit trials. The coarse resolution of the temporal representation of the task events that we introduced in the model (see Figure 3.5b and Methods) allows us to test this conclusion. To analyze its action on the RPE, we have aligned the simulated hit trials at the onset of the stimulus and confirmed that the limited temporal resolution does generate a negative modulation of the tonic activity that starts about half a second immediately before the go cue (cyan line in Figure 3.8a).

The model also predicts a decreasing activity in all the other types of trials (Figure 3.8b). In CR trials both the coarse resolution of the temporal representation and the variability in the duration of the interval between the key down event and the go cue could contribute to this effect. Since this variability spans a 2 s-interval, the decay is expected to start about 2 s immediately before the go cue. To check this in the model, we aligned simulated CR trials at the key down event and averaged them by keeping each trial only until the time when the go cue was presented. The resulting quantity exhibits the expected decay (Figure 3.8b, middle). Since the precision of this timing is affected by the limited resolution of the temporal representation, this signal starts decreasing slightly sooner. The effect is weak, but it is apparent in the traces in Figure 3.8b.

According to the model, most FA trials (84%, see Figure 3.6e) arise from transient events that occur at random times during the possible stimulation window (right panel of Figure 3.6d). Since FA events behave as low amplitude true stimuli, they generate an expectation of the go cue roughly 3.5 s immediately after their time occurrence. Therefore, they produce a slow negative modulation in the RPE beginning approximately 5 s after the key down event (because the first possible production time of FA events is around 1.5 s after the key down event), as shown in Figure 3.8b, top.

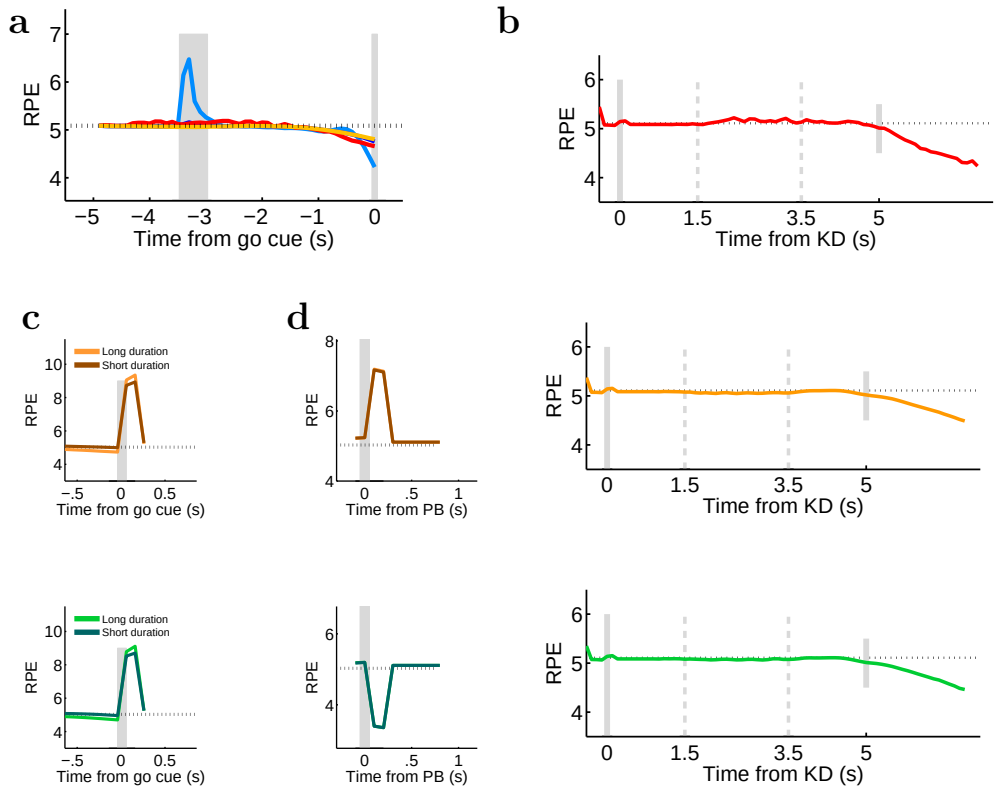


Figure 3.8: **The RPE is modulated by temporal expectation only before the go cue delivery.** (**a**) In the four trial types before the go cue the RPE showed a slow negative modulation similar to that observed in the data (Figure 3.1b). The decreasing tonic activity was particularly evident in hit trials where it was generated by the finite resolution of the temporal representation used in the model. (**b**) Top: The RPE in FA trials aligned with the key down event (KD). Note the slight positive modulation inside the possible stimulation window. Middle: The RPE aligned to KD for CR trials. In both the data and the model the activity started to decrease around the first time when the go cue could appear (gray short line on the right). This negative deviation from the baseline seemed to code a form of negative RPE that became more pronounced as the time elapsed and the expectation of the go instruction increased. Bottom: Same as in the top panel but for low amplitude miss trials. Given the lack of response to the stimulus presentation in this fraction of miss trials both the RPE and the DA activity anticipated the go cue presentation similarly to what was observed in CR trials. (**c**) When CR trials were sorted according to the duration of the key down event-go cue interval the RPE at the go cue resulted stronger for long duration trials (top). The same effect is seen for miss trials of low amplitude (bottom). The gray lines in **a**,**b** indicate the go cue. (**d**) The response of neurons to the reward delivery was independent of the trial duration in CR trials (top) and in miss trials of low amplitude (bottom). The gray lines in **c**,**d** indicate the push button event (PB).

Also, notice the slight elevation of the RPE during the window of possible stimulation, as a consequence of the random production times of the FA events (as described in Figure 3.6d). Similar effects are seen in the data (Figure 3.4c, bottom).

To complete the study of temporal expectation in the detection task, we now come back to the analysis of the dependence on the duration of the trial of the phasic response to the go cue. As we have already seen the largest DA firing activity occur for long durations (Figure 3.4a). The same behavior is seen in the model simulations in both CR (Figure 3.8c, top) and low amplitude miss trials (Figure 3.8c, bottom). It could be argued that long trials should produce a response smaller than short ones because the longer the interval, the higher the hazard for the occurrence of the go cue and the better its prediction by the RL module (Fiorillo et al., 2008). However, the response to the go cue is also affected by the finite resolution of the temporal representation. Longer intervals are represented more coarsely than short intervals and the occurrence of the go cue becomes more difficult to predict in these trials. Hence, for some value of the temporal resolution, the response to the go cue becomes larger for long intervals than for short intervals. Again in agreement with the data, where the DA phasic activation at reward delivery does not depend on trial duration (Figure 3.4b), the RPE after that event is the same for long-duration trials and short-duration trials. This result is shown in Figure 3.8d top and Figure 3.8d bottom, for CR trials and low amplitude miss trials, respectively.

Summarizing, during the variable interval in the task the RPE is modulated by the hazard function for the occurrence of the go signal. The limited resolution in the estimation of time intervals produces a similar modulation in hit trials. The hazard function, together with the imprecise temporal estimation, determines the phasic response to the go cue.

3.3 Discussion

Perceptual decisions under uncertain conditions cannot be based on the true state of the environment. Rather, noisy observations have to be combined with an internal estimate of the state, referred to as the belief state. This is the basic scheme followed in early proposals about how to extend the RL framework to model the DA activity in decision-making tasks (Dayan and Daw, 2008; Rao, 2010; Bogacz and Larsen, 2011). In this approach, the belief state is used to predict rewards, to compute the error in the prediction and to select the action that indicates the final choice. On the experimental side, de Lafuente and Romo,

2011 studied a detection task in which in each trial the animal made a choice about the presence or absence of a vibrotactile stimulus. Their main finding was that the response of midbrain neurons to a go signal reflected an internal process that they termed decision certainty, that is, the certainty the animal had on its choice. Here, to investigate this and related issues further, we adopted a different approach that allowed us to identify the type of certainty coded by the go signal and to elucidate the reasons why this certainty becomes visible at that task event. To achieve this, we defined a RL model based on the belief about the presence of the stimulus and three other features, suggested by our empirical observations: the transmission of transient activity events from a Bayesian module to a RL module; the salience of the task events and a temporal representation of those events with limited resolution.

Although other authors have included belief states (Rao, 2010), reset mechanisms (Suri and Schultz, 1999) and temporal finite resolution (Ludvig et al., 2008) separately in RL models, the need to consider them together in tasks with uncertain reward-predicting stimuli has not been noticed before. Transient increases in the firing rates of DA neurons appear in hit trials, high amplitude miss trials at the onset of the stimulus and plausibly in FA trials during a possible stimulation window. In the model, the strength of these transient events conveys the belief (and certainty) about the presence of the vibrotactile stimulus. This certainty remains hidden during the period preceding the go cue but it becomes evident in the response to this signal, generating a graduation of the RPE according to the trial type. This visibility is due to two robust properties of the model: (i) transmitted transient events of higher strength predict a higher reward (Figure 3.7a); (ii) due of the salience of the go signal the predicted reward after the delivery of the go cue is roughly independent of the occurrence of a transient event and on its strength (Figure 3.7a,b). As a consequence, the RPE is smallest for transient events of large strength and is largest in the absence of those events. This means that the RPE is large in CR trials, slightly smaller in miss trials and takes its smallest values in FA trials followed by hit trials (Figure 3.7C), in agreement with the graded DA phasic response observed in the data (Figure 3.1b). Our results help to clarify up to which point the DA response to the vibrotactile stimulus correlates with its perception. The uncertainties about the presence or absence of the vibrotactile stimulus and the time when it is applied cause a trial type dependent activity in DA neurons during the possible stimulation window. Part of this variability comes from a detection process occurring in a Bayesian module. Detection of a true stimulus produces a transient response in a RL module and leads to hit trials. Non-detected stimuli lead to miss trials. A perhaps less expected phenomenon

is that in high amplitude miss trials, the stimulus is detected, but the variability of the action selection process produces a stimulus-absent choice. In these trials, the model predicts that the cortical detection of the stimulus activates the DA neurons, although the animal's report indicates that it did not perceive it. More interesting is the case of FA trials. The average of the firing rate of DA neurons over these trials exhibits a positive modulation throughout the interval of possible stimulation. A modulation is not apparent in CR trials, although in both trial types the stimulus was not presented. The explanation comes from a recent study on cortical premotor neurons (Carnevale et al., 2015), that found that FA trials arose from transient activity events similar to those evoked by low amplitude stimuli. Consistent with this finding, our study found that the positive modulation observed in the DA activity might arise from transient cortical inputs produced at random times within the period when the stimulus is expected. In conclusion, perception is normally accompanied by a transient increase of the DA activity during the possible stimulation window, except in high amplitude miss trials. In this case, although the stimulus induced a response of the DA neurons the animal indicated that it did not perceive it.

Interestingly, the DA activity during the period preceding the go cue codes temporal expectation. The mean firing rate starts to deviate from its baseline value around the first time when the go cue can appear. As time elapses, the deviation increases in magnitude resembling a form of negative RPE strictly related with temporal expectation of the forthcoming cue. In addition to this negative slow modulation, we found that also the DA phasic activation at the go cue depends on the duration of the temporal interval preceding it, resulting stronger for long intervals. While some previous results appear not to be in contradiction with this pattern of phasic activation (Bromberg-Martin et al., 2010), other studies (Fiorillo et al., 2008; Nomoto et al., 2010) reported an opposite trend (stronger response for short intervals). Here we propose an explanation of this discrepancy: the size of the response to the go cue is determined by the hazard of occurrence of this event and by the finite resolution in the estimation of the elapsed time, which is worse for long (as in our work) than for short intervals (as, e.g, in Fiorillo et al., 2008). This explanation is consistent with an argument made in a somewhat different investigation: in a contextual instrumental task in which the hazard of occurrence of a rewarded cue increased with the number of trials elapsed since its previous appearance, Nakahara et al., 2004 found that during the early stage of learning, the response to this cue did not decrease with that number. It was argued to be due to the large counting errors produced during that stage. As in our detection task, the different responses after long or short intervals are due to

the limited resolution in the estimation of time.

In our model, task events initiate an internal representation with coarse temporal resolution. Recent works have provided direct evidence for a representation of time in the striatum that is distributed over a set of neurons (Adler et al., 2012; Mello et al., 2015). The specific set of functions adopted in our work (Shankar and Howard, 2012; Tank and Hopfield, 1987) is a possible realization of these findings. Although there are alternative temporal representations (Ludvig et al., 2008) and approaches (Daw et al., 2006), our choice was dictated for the sake of simplicity and because there exist detailed studies of this internal representation that makes its use attractive (Shankar and Howard, 2013; Shankar, 2015).

The main results of the model rely on robust features that depend little on the precise parameter values. Partly for this reason and also because of the difficulty of the computation, we did not attempt to fit the model to the DA electrophysiological data. Instead we preferred to identify the mechanisms that can explain how the DA activity is modulated by the stimulus and temporal uncertainties present in the task. In addition, some parameters have been set according to physiological constraints; this is the case of the input noise, in the model it appears as Poisson spike trains with firing rates set to values similar to those observed in prefrontal neurons. Hence, the events transmitted to the RL module were not controlled by tuning the input noise. Other features of the model did not required new parameters; for instance the reset induced by the salience of the task events is based on the direct comparison between the reward predicted by the current event and that predicted by the events preceding it, without including any specific threshold parameter. The parameters associated with the limited resolution of the temporal representation of the task events were set in such a way that the decay of the RPE at the end of the interval preceding the go signal was similar to the decay observed in the data. The same values of those parameters yield a dependence of the phasic response to the go signal on the duration of the trial larger for the longest trial durations. The discount factor, also relevant to describe this phenomenon, was fixed at $\gamma = 0.98$, which is a standard value for this parameter.

Research about learning by reinforcement when reward predicting stimuli are uncertain could be extended in several directions. Here we followed a normative approach; but in reality learning occurs in networks of spiking neurons. Although there have been recent promising advances in training this type of networks, the methods are based on supervised learning (Abbott et al., 2016) instead of on reinforcement learning (but see Friedrich and Lengyel, 2016; Friedrich et al., 2011; Vasilaki et al., 2009; Potjans et al., 2009). Assess-

ing the generality of our conclusions would require consideration of other experimental paradigms to guide the search for relevant features to be included in more complete RL models. An intriguing case is the discrimination between two sequential stimuli, when some physical property of one of them has to be kept in working memory before the presentation of the second one. An example of this is the somatosensory discrimination task thoroughly studied in several cortical areas (Romo et al., 1999; Hernández et al., 2010). This requires including memory in the model (Kaelbling et al., 1998; Todd et al., 2009) and dealing with an uncertainty that does not reside in the stimuli themselves, but in the comparison between the second stimulus and the memory of the first. In the purely temporal domain, the study of tasks that compare two temporal patterns of pulse stimuli (Rossi-Pool et al., 2016) would help to define the most convenient temporal representations. A systematic study of these and other paradigms often used to investigate decision making processes would contribute to understand how DA influences the learning of associations between stimulus and reward under uncertain conditions.

The results obtained in the model and experimental data show that the RPE signal codes also (i) the animal's certainty about the presence of the stimulus; (ii) the temporal expectation of reward predicting sensory cues; (iii) to some extent, also the perception of uncertain stimulus. As it is proposed by the model, these processes take place in a Bayesian (plausibly cortical) module, which are then sent to a RL module (plausibly the midbrain DA system and the striatum). The results of the model and the experimental data show that the activity of the DA neurons is not a mere reflection of the cortical signals but rather that they are transformed into a new signal with a quite different function. However, some expressions of the original inputs are still visible in the firing rate of the DA neurons. These are, for example, the transient events that are related to decision making processes; the certainty about the presence of these events originates a hierarchy of responses to cues predicting reward; and the acquired knowledge about the stochastic temporal structure of the trials produces a declining DA activity during the intervals between task events. Whether these processes depend only on the inputs to the DA neurons and the RL computations performed over them or if there is further elaboration in the midbrain DA system is an open question.

3.4 Methods

Detection Task. Monkeys were trained to detect a vibrotactile stimulus of variable amplitude applied to one of its fingertips (de Lafuente and Romo, 2005). Stimulus-present

trials were randomly interleaved with an equal number of stimulus-absent trials. Stimuli were delivered to the skin of the distal segment of one digit of the restrained hand, via a computer-controlled stimulator (BME Systems; 2 mm round tip). Initial probe indentation was 500 μm. Vibrotactile stimuli consisted of trains of 20 Hz mechanical sinusoids with 9 different amplitudes between 2.3 and 34.6 μm. Crucially some of the amplitudes were very weak and consequently difficult to detect. Animals were rewarded with a drop of liquid for correct behavioral responses (correct detections in stimulus-present trials and correct rejections in stimulus-absent trials) and received no reward otherwise (miss trials and false alarm trials). Animals were handled in accordance with standards of the National Institutes of Health and Society for Neuroscience. All protocols were approved by the Institutional Animal Care and Use Committee of the Instituto de Fisiologa Celular.

Recordings. Data for this analysis were obtained from an earlier study (de Lafuente and Romo, 2011). Recordings were obtained with quartz-coated platinum-tungsten microelectrodes (2 to 3 MΩ; Thomas Recording) inserted through a recording chamber located over the central sulcus, parallel to the midline. Midbrain DA neurons were identified on the basis of their characteristic regular and low tonic firing rates (1-10 spikes per second) and by their long extracellular spike potential (2.4 ms $\pm$ 0.4 SD). Among the 69 neurons analysed in the previous work we selected a group of 23 cells. The selected group of cells corresponded to those neurons which response to the reward delivery did not violate reinforcement learning principles: they showed a positive phasic activation or lack of response in correct trials (hit and correct rejection trials) while the activity paused or remained at the baseline level when the reward is omitted (miss and false alarms trials). A similar criterion has been adopted in many electrophysiological studies of midbrain dopamine neurons (Morris et al., 2004; Roesch et al., 2007; Takahashi et al., 2016).

Data Analysis. For each neuron, we computed the firing rate as a function of time using 300 ms sliding windows displaced every 50 ms (Figure 3.1b). Responses to the stimulus (in Figure 3.1c and in Figure 3.2a, right) were measured in a 500 ms window centered 350 ms after the stimulus onset and were standardized with respect to a pre-stimulation window (of 500 ms centered 700 ms before the stimulus presentation). Responses to the go instruction (Figure 3.3, left) were measured in a 250 ms window centered 170 ms after the instruction and were standardized with respect to a precue window (of 250 ms centered 500 ms before the cue presentation). Responses to the reward delivery were measured

in a 400 ms window centered 350 ms after the push button and were standardized with respect to a precue window of 200 ms centered 200 ms before the push button (Figure 3.3, right). The activity outside the possible stimulation window (PSW) was calculated in two 1 s windows before the start and after the end of the PSW (from 500 ms to 1.5 s after the key down event and from 3.7 s to 4.7 s after that event). The mean activity during and outside the PSW was standardized with respect to a 500 ms windows centered 1 s after the key down event (Figure 3.2b).

Model. The model relies on two modules: a bayesian module and a reinforcement learning (RL) module. The first module uses observations to estimate a posterior probability (belief) about the current state of the external world s_t. More specifically it calculates the belief $b_{sp}(t)$ about the presence of the (ambiguous) vibrotactile stimulus:

$$b_{sp}(t) = P(s_t = sp|X_{1:t}) \tag{3.1}$$

where $X_{1:t}$ is the entire history of observations up to time t. The stimulus is assumed to be detected by the bayesian module when $b_{sp}(t) > 0.5$ (i.e using the *maximum a posteriori* -MAP- criterion). The latter module consists in a standard RL architecture known as Actor/Critic (Barto, 1995). We consider a total of 6 events: the vibrotactile stimulus, the start and go signals, the response movements of the animal (key down and the 2 push buttons indicating yes/no responses).

The physical salience function of event i is represented by the i-th component of the vector $\mathbf{e}(t)$. With the exception of the vibrotactile stimulus, the component $e_i(t)$ takes value one at the onset of the event i and zero otherwise. The component $e_v(t)$ corresponding to the vibrotactile stimulus is activated when the bayesian module detects it. In this case we set $e_v(t_d) = b_{sp}(t_d)$ (with t_d denoting to the time of the detection).

The onset of the salience function $e_i(t)$, at time t^i_{on}, activates a temporal representation $\mathbf{x}_i(t)$ of the event i. This is defined as a set of N functions $T_{i,m}(t)$ $(m = 1, \ldots, N)$, each representing the event (a pulse of one time step duration) around time τ_m after its detection. We assume that the resolution of these functions decreases with τ_m and that the times τ_m are distributed uniformly on a logarithmic time scale (from a minimum value $\tau_{\min} = 0.1s$ to a maximum value $\tau_{\max} = 10s$). This leads to a scale invariant representation of the event i. An explicit mathematical realization is (Shankar and Howard, 2013):

$$T_{im}(t) \equiv T_i(t - t^i_{on}, \tau_m) = \frac{1}{|\tau_m|} C(k) \int_{d_i(t)}^{a_i(t)} \left(\frac{\tau'}{\tau_m}\right)^k e^{-k\frac{\tau'}{\tau_m}} \, d\tau' \tag{3.2}$$

where $C(k) = k^{k+1}/k!$, $a_i(t) = t^i_{on} - t$, $d_i(t) = t^i_{on} + dt - t$ and dt is the duration of the original pulse (alternatively Equation 3.2 could be expressed as a convolution of an

alpha function with a pulse). The parameter k controls the smear in the representation (the larger is k the more accurate is the representation). The temporal representation $\mathbf{x}_i(t) = \{x_{i1}(t), x_{i2}(t), \dots x_{lN}(t)\}$ is taken equal to the functions in Equation 3.2 multiplied by the physical salience function of the event i:

$$\mathbf{x}_i(t) = e_i(t^i_{on})\mathbf{T}_i(t) \tag{3.3}$$

The reward predicted by the event i is expressed as:

$$P_i(t) = \sum_{m=1}^{N} x_{im}(t) w_{im} \tag{3.4}$$

The total predicted reward at time t, $V(t)$, is given by:

$$V(t) = \sum_i P_i(t) \tag{3.5}$$

Following (Suri and Schultz, 1999), we suppose that the occurrence of an event i with reward prediction higher than the total reward prediction at previous time disrupts earlier events representations:

$$P_i(t^i_{on}) > V(t^i_{on} - 1)/\gamma \implies x_{j,m} = 0, \quad j \neq i \tag{3.6}$$

The DA signal is assumed to be represented by the reward prediction error RPE. However, DA neurons show an asymmetrical activity due to their low baseline firing rate. This asymmetry is taken into account by introducing a rectification threshold $\psi > 0$ for the RPE:

$$\delta(t) = \begin{cases} r(t) + TD(t) & \text{if} \quad r(t) + TD(t) > -\psi \\ -\psi & \text{otherwise} \end{cases} \tag{3.7}$$

where $TD(t) = \gamma V(t) - V(t-1)$ and $r(t)$ takes the value R if the reward occurs at time t and 0 otherwise. The ratio between the value of ψ and the scalar reward value R determined the degree of asymmetry in the error signal (the asymmetry increases if the ratio decreases).

The weights w_{im} in Equation 3.4 are adapted during learning as follow:

$$\Delta w_{im}(t) = \begin{cases} \eta^+_c x_{im}(t)\delta(t) & \text{if} \quad \delta(t) > 0 \\ \eta^-_c x_{im}(t)\delta(t) & \text{if} \quad \delta(t) < 0 \end{cases} \tag{3.8}$$

where η^+_c indicates the learning rate for acquisition and η^-_c is the learning rate in extinction.

The input to the actor component is a vector trace $\bar{\mathbf{e}}(t)$, which components $\bar{e}_i$ are defined as:

$$\bar{e}_i(t) = e_i(t) + \rho\, \bar{e}_i(t-1) \tag{3.9}$$

where $\rho < 1$ is a decay parameter.

The actor selects an action a_j only at the end of each trial, after the go cue. The possible actions are pressing one of the two buttons corresponding to yes/no decisions (the action of withholding movement is not allowed). The probability of choosing the action a_j for an input $\bar{\mathbf{e}}(t)$ is given by a softmax distribution:

$$P(a_j|\bar{\mathbf{e}}(t)) = \frac{e^{\sum_i \nu_{ji}\bar{e}_i/\beta}}{Z} \tag{3.10}$$

where Z is the normalization constant and the parameter β governs the exploration/exploitation trade-off: as β approaches 0, action selection approaches a winner-take-all mode while larger values of β favour exploration.

The weights ν_{ji} in Equation 3.10 are adapted only at the end of each trial when the reward is expected. Pressing of one of the two buttons occurs 3 time steps (i.e 0.3 s) after the go cue. The reward is delivered 2 time steps after the movement. The weights ν_{lj} are adapted with the following learning rule:

$$\Delta\nu_{ij} = \begin{cases} \eta_a^+ \sum_t \bar{e}_i(t_r)\delta(t) & \text{if} \quad j = \bar{j}, \delta(t) > 0 \\ \eta_a^- \sum_t \bar{e}_i(t_r)\delta(t) & \text{if} \quad j = \bar{j}, \delta(t) < 0 \\ 0 & \text{if} \quad j \neq \bar{j} \end{cases} \tag{3.11}$$

where $\bar{j}$ denotes the selected action and t_r is the time when the reward is expected (i.e 5 time step after the go cue). The parameters η_a^+ and η_a^- correspond to the learning rate in acquisition and in extinction.

Model Analysis. In all the simulations we used a time bin $dt = 100$ms (for a full list of parameters used in the model see **??**). To compare the model results with the mean activity of DA neurons we transformed the simulated RPE $\delta(t)$ in an equivalent firing rate $[\delta(t)]_{\text{equiv}}$ as follows:

$$[\delta(t)]_{\text{equiv}} = baseline + F * \delta(t) \tag{3.12}$$

The *baseline* representing the baseline activity of DA neurons during the trial was set to 5.1 Hz. The value of the scale factor F was chosen to obtain an equivalent prediction

error $[\delta(t)]_{\text{equiv}}$ that matched the mean dopamine response at the start cue. Its value in all the simulations was 27.5 Hz. Additionally, the signal $[\delta(t)]_{\text{equiv}}$ was filtered using a 300 ms sliding window displaced every 100 ms (a procedure equivalent to the one done to obtain the firing rate of DA neurons as a function of time). Responses to the stimulus (in Figure 3.6b) were calculated averaging the signal $[\delta(t)]_{\text{equiv}}$ over a 300 ms window centered 100 ms after the stimulus onset. Responses to the go instruction and to the reward delivery were calculated averaging the signal $[\delta(t)]_{\text{equiv}}$ over a 300 ms window centered, respectively, 100 ms after the go cue and after the reward delivery (Figure 3.6a).

	Description	Symbol	Value
Critic	Learning rate in acquisition	η_c^+	0.1
	Learning rate in extinction	η_c^-	0.2
	Rectification	ψ	0.15
	Discount Factor	γ	0.98
	Smear of the **T** functions	k	40
	Spacing of the **T** functions	c	0.2
Actor	Learning rate in acquisition	η_a^+	0.03
	Learning rate in extinction	η_a^-	0.1
	Noise of the softmax	β	0.5
	Decay of stimulus trace	ρ	0.98

Chapter 4

The dopamine signal in tasks involving parametric working memory

4.1 Introduction

Working memory refers to the ability to hold information available for processing during short periods of time (Baddeley and Hitch, 1974). Evidence for storage of working memory contents in multiple brain regions has been provided by many studies in humans and non-human primates (Christophel et al., 2017). In particular, it is almost universally accepted that the prefrontal cortex (PFC) plays a critical role in the dynamical control of items stored for short periods of time. Furthermore, it has long been known that the dopamine (DA) system interacts closely with the PFC (see for example Alexander et al., 1986) and that damage to the DA system can impair cognitive functions typically associated associated with this region (Moghaddam et al., 1997).

In task involving working memory manipulations of the DA receptors crucially affect the behavioural performance (Sawaguchi and Goldman-Rakic, 1991; Arnsten et al., 1994; Murphy et al., 1996) and single cell recording provided evidence for effects of DA on PFC activity during task performance (Williams and Goldman-Rakic, 1995; Sawaguchi, 2001). It has been suggested that the activity of DA neurons affects working memory in a dual way: high tonic DA activity are thought to be relevant for maintaining information in PFC, while phasic burst of DA neurons allows rapid updating and learning (Cohen et al., 2002). However the role of DA neuron during tasks involving working memory is yet to be clarified.

To get more insights into this issue we focus our attention on the DA signal recorded from behaving monkeys performing a well studied example of perceptual decision making requiring working memory, the discrimination task reported in (Romo et al., 1999; Romo and Salinas, 2003; see Figure 4.1). This task consists in the sequential comparison of two vibrational stimuli separated by a delay period of a few seconds.

Extensive analysis of single cell activity revealed neural correlates of working memory in multiple brain areas during the delay period between the first (base) stimulus and the second (comparison) stimulus. In general, except for primary sensory cortex (S1), which activity did not to show history-dependence (i.e dependence on the base frequency) neither during the delay not at the presentation of the comparison stimulus, neurons in other cortical areas showed activity correlated to stimulus value, memory, and decision outcome (see for example Romo et al., 1999; Hernández et al., 2002, 2010).

Of particular interest here is the activity in PFC, for the well known interaction between PFC and the DA system in working memory tasks mentioned above. Previous reports of the neural activity in this area showed that the the firing rate of many of the cells either increases or decreases monotonically with the base stimulus frequency. The monotonic tuning was observed not only during the stimulus presentation but also at different time points during the delay period after the stimulus offset (Romo et al., 1999). When the interval of the delay duration changed neurons rescaled their time-dependent firing rate according to the new interval duration (Brody et al., 2003). Importantly, the subpopulation of neurons that are tuned to the stimulus was not constant in time. This subpopulation decreased during approximately the first second after the base stimulus presentation and became more and more prominent as time elapsed through the end of the delay (Romo et al., 1999; Jun et al., 2010). A subsequent analysis (Barak et al., 2010) seemed indeed to support the idea that the quality of stimulus encoding by the population deteriorated after the offset of the base stimulus, and gradually recovers toward the end the delay period.

In what follows we present new data of the DA activity recorded during the discrimination task. A model based analysis of these new experimental results supports the hypothesis that the phasic DA reward prediction errors (RPEs) are shaped by a non-trivial probabilistic inference sensory processing. Surprisingly, data showed a high tonic positive modulation of the DA activity during the delay period. According to the model also this sustained tonic modulation can be interpreted as a form of RPE relying on an inference process, likely to take place in prefrontal areas, and transmitted to the DA system during the delay period.

4.2 Results

4.2.1 Discrimination Task and Behavioral Performance

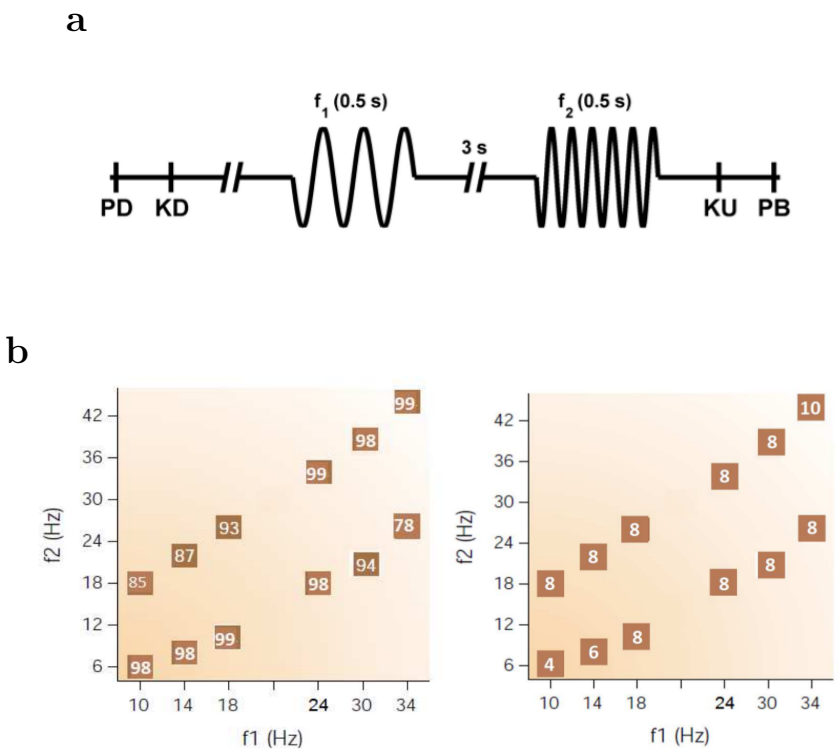


Figure 4.1: **Discrimination task.** (**a**) Trials began when the mechanical probe is lowered (probe down, PD). The monkey reacted by placing its free hand on an immovable key (key down, KD). After a variable period (1.5-3.0 s) the probe oscillated for 0.5 s at the base (f_1) frequency. A second vibrotactile stimulus at the comparison frequency (f_2) was presented 3 s after the offset of the first stimulus. At the end of the second stimulus the monkey released the key (key up, KU) and pressed one of two push-buttons (PB) to indicate whether the comparison frequency was higher or lower than the base. (**b**) Stimulus sets (i.e the pairs of f_1 , f_2 used in the task. Numbers inside the brown boxes indicate percentage of correct trials in each condition (left), and the absolute value of the difference between the base and the comparison frequency (right).

During the vibrotactile discrimination task (Romo et al., 1999; Romo and Salinas, 2003), the monkey had to pay attention to the frequency of the base stimulus, remember this frequency for a delay of 3 seconds, and then compare it to the frequency of the second comparison stimulus. The time course of an individual trial is shown in Figure 4.1a. The

animal obtained reward for correctly identifying the higher frequency. The sets, the pair of frequencies, used during the recordings are illustrated in in Figure 4.1b (left) together with the absolute value of the difference between the base and comparison frequency.

4.2.2 Phasic DA response to the first vibrotactile stimulus

At a population level, the onset of the base stimulus positively activated the DA neurons (see Figure 4.3a ; $p < 0.01$, Wilcoxon signed-rank test). This result indicates that DA neurons were excited by the stimulus to be retained in working memory. An encoding of the frequency has been observed in many areas, including PFC (Romo et al., 1999; Hernández et al., 2002, 2010). We therefore asked whether the mean activity of the population encoded the specific frequency value that needed to be stored. Figure 4.2a shows the response of the neurons to the first stimulus sorted by the value of the base frequency. According to this analysis the population activity did not show any significant dependence on the value of the first stimulus frequency (one-way ANOVA, $p = 0.26$). However, data exhibited a slight tendency for a stronger activation for stimulus frequencies toward the center of its value distribution. We analysed therefore whether this pattern of activation could be related to the encoding of some reward-related information. Naively, since the trial condition is not fully defined until the application of the second stimulus, the phasic DA response to the first stimulus should not depend on the value of f_1. However this cannot be correct, as an inspection of the fraction of correct responses for each pair (f_1, f_2) indicated. In fact, for the stimulation set used in the experiment, the performance at fixed f_1 was worse at the two end values of this frequency (Figure 4.1b, left). This anomaly in the performance appears in delayed comparison tasks (first noticed in Hollingworth, 1910), and it is known as the contraction bias. The bias consists in judging the magnitude of the first stimulus as greater than that of the second one, when the two stimuli are small in magnitude. The opposite happens for two the stimuli that are relatively large. So, intuitively, it can be described as a shift of the perceived frequency of the first stimulus to the center of its range. For example, if f_1 is selected at the lower end of its range, it will be judged as being larger. Then one expects that for the pair ($f_1 = 10$ Hz, $f_2 = 18$ Hz) the performance would deteriorate with respect to that for values of f_1 taking values far from the ends of its range, as it is the case (Figure 4.1b, left). The same is true for the pair ($f_1 = 34$ Hz, $f_2 = 18$ Hz). This phenomenon suggests that the higher DA responses toward the center of the f_1 distribution could be due to different predictions of reward.

To check the relevance of this reward-related effect we have compared the firing rate of the DA neurons during the application of the first stimulus of the two values of f_1

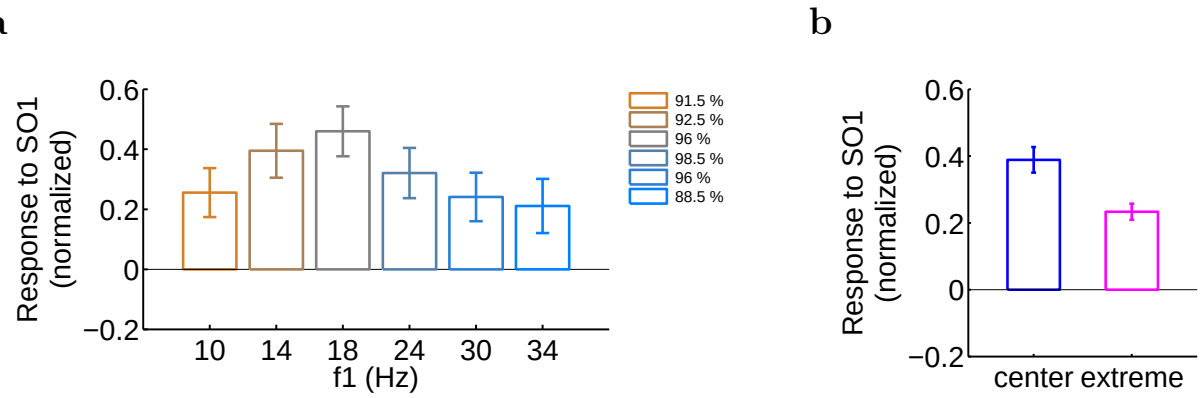


Figure 4.2: **DA response to the first stimulus.** (**a**) Population response (z-score) to the onset of the base stimulus sorted by the frequency of the vibration. (**b**) Population response (z-score) to the onset of the base stimulus for 'extreme' frequencies (i.e. $f_1 = 10$ Hz and $f_1 = 34$ Hz) and for 'center' frequencies (i.e $f_1 = 18$ Hz and $f_1 = 24$ Hz). Error bar are $\pm$ 1 SEM.

situated at the extreme of the its possible values ($f_1 = 10$ Hz and $f_1 = 34$ Hz, denoted as extreme) with the two central values ($f_1 = 18$ Hz and $f_1 = 24$ Hz, denotes as center). Although the difference in activation did not reach significance value ($p = 0.067$, two tail t-test), the result of this analysis showed a clear tendency for larger activation when the base stimulus frequency corresponded to trial types that resulted easier for the animal, and thus more likely to be rewarded (see Figure 4.2b).

4.2.3 Modulation of the DA activity during the delay period

Individual neurons in several prefrontal areas exhibit persistent activity during the delay period tuned parametrically to the value of the frequency of the first stimulus (Romo et al., 1999; Hernández et al., 2010). Besides, the traces of their firing rates are heterogeneous and not uniform in time (Brody et al., 2003). This poses some intriguing questions about how the activity of DA neurons behaves during the delay period. Are they temporally modulated during that period? Do they code the value of f1? Do they code some reward-related information? To answer the first question, for each neuron we have averaged its spiking activity over all trials. In Figure 4.3c we present the result of this computation for two example neurons. Note that the firing rate of the first neuron is clearly modulated in time, increasing throughout the duration of the delay period (Figure 4.3c, left). Instead, the firing activity of the other neuron has a more uniform temporal behaviour (Figure 4.3c,

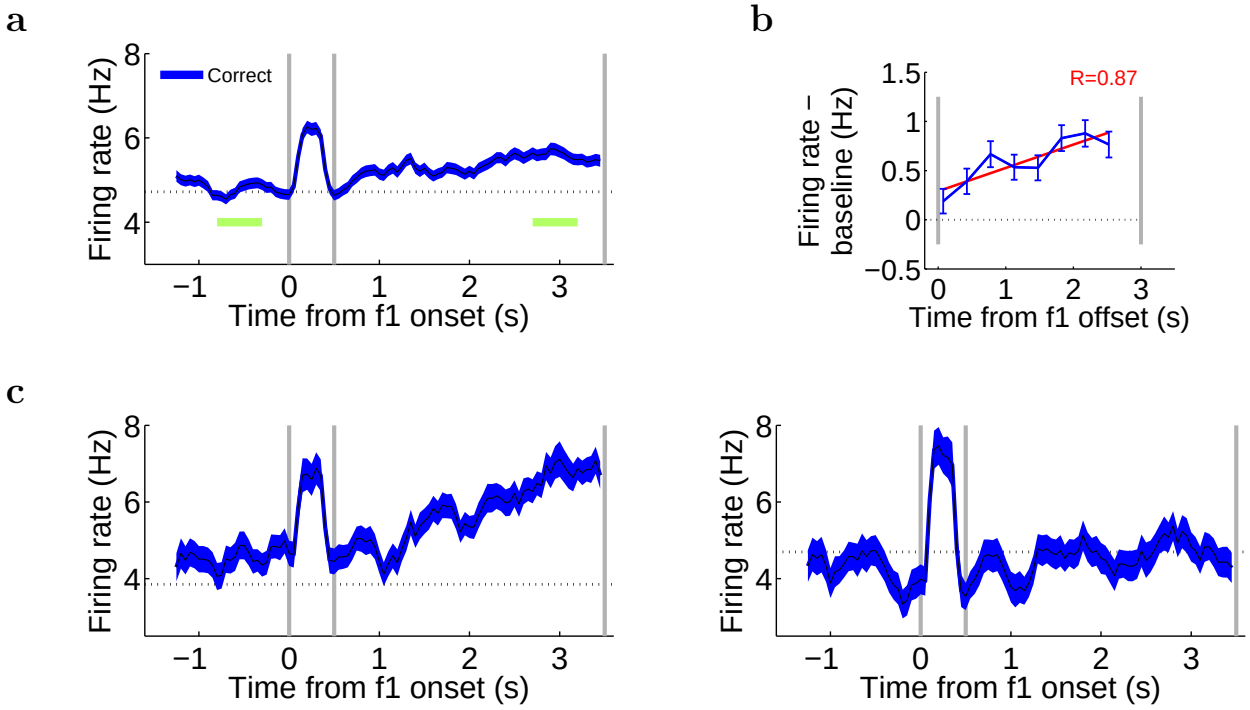


Figure 4.3: **DA activity during the delay period.** (**a**) Mean population firing rate (black line, ± SEM colored band) plotted as a function of time for correct (rewarded) trials. Trials are aligned to the onset of the base stimulus. The gray bars (from left to right) indicate the onset and the offset of the first stimulus, and the onset of the second stimulus. (**b**) Average activity of DA neurons during the delay period. Gray vertical lines mark the offset of the base stimulus and the onset of the comparison stimulus. Blue lines are mean baseline-subtracted firing rate in non-overlapping bins of 350 ms. Error bars are ± 1 SEM. (**c**) Same as in **a** but for two example neurons.

right).

At a population level DA neurons show clear positive tonic modulation during the delay period. The average firing rate starts to increase immediately after the offset of the first stimulus and does so until the presentation of the second stimulus (Figure 4.3a-b). Indeed the mean activity of the population before the presentation of the first stimulus is significantly lower than the mean activity before the presentation of the second stimulus ($p < 0.01$, Wilcoxon signed-rank test; both means are calculated in a 500 ms which position is indicated by the green horizontal lines in Figure 4.3a).

We next tried to determine the origin of this positive modulation in the delay activity. It is known that the fraction of selective neurons in prefrontal areas is not temporally

homogeneous (Romo et al., 1999; Hernández et al., 2010). The size of this subpopulation decreases during approximately the first second after the base stimulus presentation and becomes more and more prominent as time elapses through the end of the delay. A subsequent analysis (Barak et al., 2010) seemed indeed to support the idea that the quality of stimulus encoding by the population deteriorates after the offset of the base stimulus, and gradually it recovers toward the end the delay period. A similar behavior of selective activity during the delay period happens in other tactile tasks where a sample stimulus has to be kept in working memory (Vergara et al., 2016; Rossi-Pool et al., 2016). We asked therefore whether the increase in the DA activity during the delay period could result from a time varying signal received from prefrontal inputs. If this hypothesis was true one would expect some tuning to f_1 during the delay period, or at least toward the end of the delay (because the activity in prefontal areas seems to show a better encoding of the base frequency immediately before the onset of the comparison stimulus).

To investigate the possibility of f_1 tuning during the delay period we computed the temporal profile of the neurons' firing rates sorting trials according to the value of the first frequency. For the same two example neurons the curves for different values of f_1 appear superimposed (Figure 4.4c). This property can also be seen clearly in the temporal averages of the firing activity (z-score) over the delay period at fixed f_1 which do not exhibit significant differences (Figure 4.4d). This absence of tuning exists also at the population level during the entire delay period (Figure 4.4a; one way ANOVA, $p = 0.63$). Even in the last 500 ms of the delay, where the activity of prefrontal areas suggest a recovery of the information about the base frequency (Barak et al., 2010), DA neurons did not exhibit any evident tuning to f_1 (one way ANOVA, $p = 0.71$).

The DA signal during the delay period is therefore quite different from the delay activity encountered in prefrontal cortex, where neurons code parametrically the memorized value of the frequency (Romo et al., 1999; Hernández et al., 2010). In principle, the lack of tuning in the frequency of the first stimulus does not contradict the functional properties of the DA neurons related to the prediction of reward. However, despite of the result obtained for the activation after the base stimulus onset, during the delay period the DA activity did not show any difference between the extreme and the central values of the f_1 distribution ($p = 0.95$, two tail t-test; compare Figure 4.4b with Figure 4.2b).

Although our analysis did not allow to determine the origin of the DA modulation in the delay activity, our data show a clear increasing DA activation from the offset of the first stimulus to the onset of the second one. This result differs from that obtained another study involving working memory. In that work Matsumoto and Takada, 2013 investigated

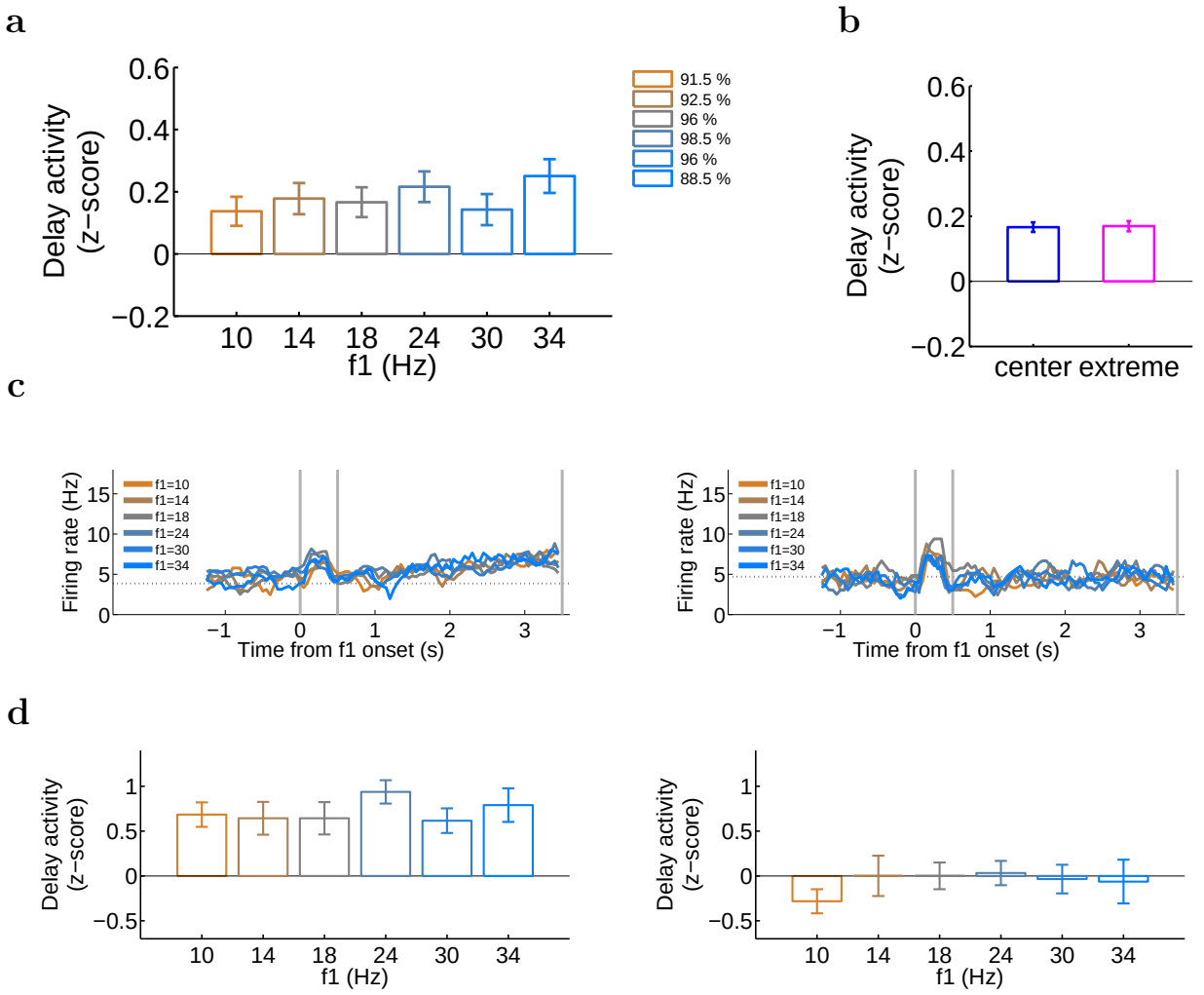


Figure 4.4: **Absence of f_1 tuning during the delay period.** (**a**) Mean normalized activity (z-score) of the population sorted according to the value of the first frequency (see Methods for details about the normalization). (**b**) Mean population response (z-score) during the delay period for 'extreme' frequencies (i.e. $f_1 = 10$ Hz and $f_1 = 34$ Hz) and for 'center' frequencies (i.e $f_1 = 18$ Hz and $f_1 = 24$ Hz). Error bar in **a** and **b** are $\pm$ 1 SEM. (**c**) Mean firing rate of the same two example neurons of Figure 4.3c sorted according to the value of the first frequency for correct trials. Trials are aligned to the onset of the base stimulus. The gray bars (from left to right) indicate the onset and the offset of the first stimulus, and the onset of the second stimulus. (**d**) Same than in **a** but for the two example neurons of panel **c**.

the DA responses in a visual search task in which monkeys had to maintain in working memory the orientation of a lighted bar for a short interval, and then discriminate the

sample stimulus within an array of bars with several orientations. In that study DA neurons did not exhibit any persistent activation during the delay period.

4.2.4 Phasic DA response to the second vibrotactile stimulus: correct and error trials

It is instructive to compare the firing response of DA neurons to the first and the second stimuli, both in correct and wrong trials. After application of the first stimulus the phasic responses in error and correct trials have a similar profile (Figure 4.5a).

In contrast, the application of the second stimulus produces a quite different behaviour of the firing rates in the two trial types. After an initial common increase, the activity in wrong trials deviates significantly from the activity in trials with correct choices (Figure 4.5b).

This difference could be related to the function of DA neurons as coding the error in the prediction of the total future reward (dopamine reward prediction errors) (see model results below). But the time when they start to depart from each other can be explained in terms of the latency of cortical signals related to inference processes about which one of the two frequencies is the largest. In fact, although in some sensory areas (i. e. S1, see Romo et al., 2002; Hernández et al., 2010) there are not significant differences between the activity in correct versus error trials, a population of neuron in S2 tuned to $(f_1 - f_2)$ distinguishes well between the two trial types after about 220 ms (Romo et al., 2002).

Neural populations tuned to the difference between the two frequencies abound in several prefrontal areas (Hernández et al., 2010). Those neurons are processing the difference of the frequencies in a way correlated with the animal's behaviour. It is then plausible that cortical neurons send this information to the DA midbrain system. In turn these neurons could use those signals to comply with their own function of computing and representing the error on the prediction of the future reward. In support of this conjecture, we notice that the responses to the two stimuli have a short latency. However, as it was noticed above, in the response to the second stimulus the activities in correct and error trials start to diverge only after a longer latency. To further quantify divergence arises at a population level, we used receiver operating characteristic (ROC) analysis and computed area under ROC curve (AUC) in sliding time windows (see section 4.4).

According to this analysis DA neurons show a latency of about 250 ms, a value which is longer than the latency of S2 neurons tuned to $(f_1 - f_2)$ and roughly equal to the latency of prefrontal neurons with similar tuning. It is then plausible that DA neurons receive

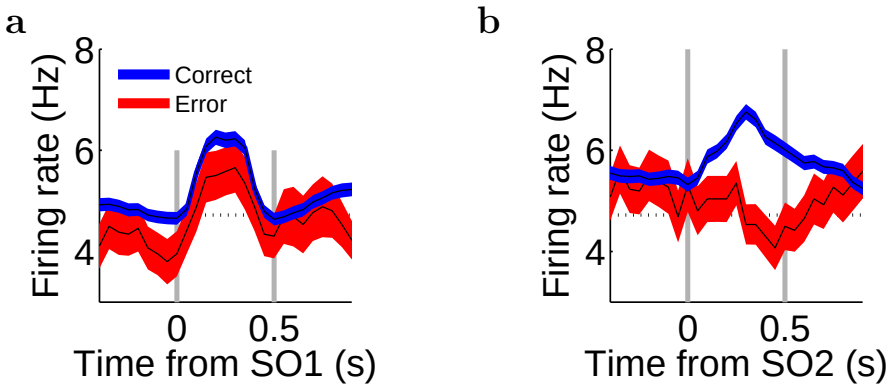


Figure 4.5: **DA response to the stimuli in correct and in error trials.** (**a**) Mean population firing rate (black line, ± SEM colored band) plotted as a function of time for correct and error trials. Trials are aligned to the onset of the base stimulus. The gray bars (from left to right) indicate the onset and the offset of the first stimulus. (**b**) Same than in **a** but with trials aligned to the onset of the comparison stimulus.

a signal coding the difference between the two frequencies. This is in agreement with previous findings, suggesting that sensory evoked DA activity does not signal the stimulus physical attributes but arises from the output of a perceptual process (de Lafuente and Romo, 2011, 2012; Nomoto et al., 2010; Lak et al., 2017). DA neurons could use that signal to compute the error in the estimate of the reward. It is therefore plausible that in correct trials the cortical inference process is more accurate than in error trials. As a consequence the quality of the perception results better in correct trials and this produces a positive dopamine RPE while in error trials a poor sensory processing of the second stimulus is responsible for a negative dopamine RPE in Figure 4.5.

4.2.5 Response of DA neurons as a function of task difficulty

How is the difficulty of the task controlled? The relevant task parameter is the difference between the frequencies of the two stimuli, $(f_1 - f_2)$. So, in principle, the task difficulty could be defined in terms of this difference. If this were correct the monkey performance in all trials with the same value of this difference should be similar. However, as discussed in subsection 4.2.2, an estimate of the fraction of correct choices shows that this is not so. Given a value of the difference $|f_1 - f_2|$ of 8 Hz, the performance for the extreme values of f_1 is rather poor (compare for example the performance of the pair of frequencies $f_1 = 24$ Hz, $f_2 = 16$ Hz, and $f_1 = 34$ Hz, $f_2 = 26$; see Figure 4.1). The behavioural anomaly is

commonly attributed to the contraction bias.

This bias was initially explained with the existence of an internal reference, which was used in the comparison tasks instead of the perceived magnitude (see for example Hellström, 1985). Recently it was proposed that the contraction bias results from a Bayesian inference computation in which noisy representations of stimuli are combined with knowledge about the a-priori distribution of magnitudes in order to optimize performance (Ashourian and Loewenstein, 2011). Intuitively, such an inference should lead to the contraction bias because the perception of extreme magnitudes of the first stimulus, which are unlikely given unimodal prior distributions, will be biased toward the 'center' of the prior distribution. The contraction bias naturally arises from Bayesian inference if one assume that the level of uncertainty (i.e the noise) in the representation of f_1 (the information that needs to be retrieved) is higher than the the level of uncertainty in the representation of f_2 (because the comparison stimulus is supposed to be present when the decision is made). This is indeed equivalent to suppose that anomaly in the performance is due to the memory retrieval/decision making process and not to the memory encoding.

In the context of our discrimination task this Bayesian contraction bias can be incorporated as follows (see section 4.4 for more details): right after the onset of the comparison stimulus the animal receives a noisy observation about the memory of the base frequency and a (less) noisy observation about the frequency of the second stimulus. These noisy informations are combined with the knowledge of the two prior distribution to calculate the posterior distributions, or beliefs, about the values of the two frequencies. Given the two beliefs about the first and the second stimulus frequency, denoted respectively as $\mathbf{b}_1(f)$ and $\mathbf{b}_2(f)$ the posterior distribution about $f_1 > f_2$ can be obtained as follows:

$$b(f_1 > f_2) = \sum_{f_i}^{f_1^{max}} b_1(f_i) \sum_{f_j<f_i} b_2(f_j) \tag{4.1}$$

The above quantity is then used by the model to make a decision about which of the two frequency is the highest (see the section section 4.4). In our analysis of the task difficulty we assume the validity of the Bayesian decision model sketched above, and adjust the level of noise in order to minimize the difference between the performance of the model and the performance of the animal(this corresponds to fit two noise parameters to the animal performance; see section 4.4 for details about the model fitting).

We then redefine the difficulty of the task according to the performance of the model. That is: easy classes are those in which the model reached the highest performance, intermediate difficult classes corresponded to those with intermediate performance in the

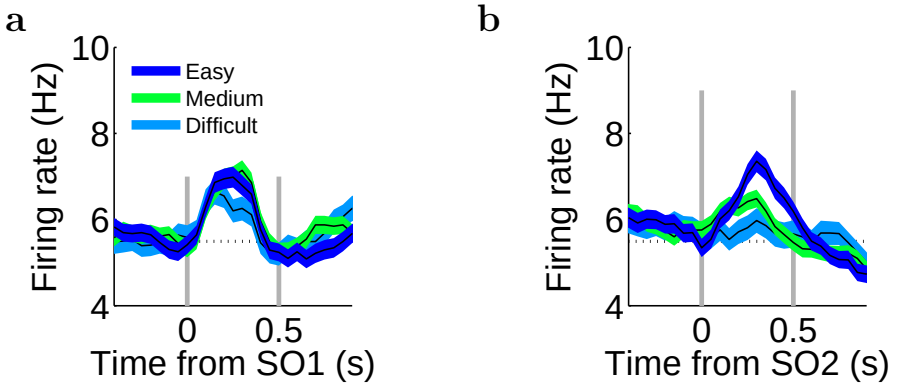


Figure 4.6: **DA response to the stimuli in correct and in error trials.** (**a**) Mean population firing rate (black line, ± SEM colored band) plotted as a function of time for correct and error trials. Trials are aligned to the onset of the base stimulus. The gray bars (from left to right) indicate the onset and the offset of the first stimulus. (**b**) Same than in **a** but with trials aligned to the onset of the comparison stimulus.

model, and difficult classes are those in which the model performed the worst. Intuitively, this way of defining the difficulty of the task, corresponds to say that the animal is indeed following an internal model similar to our Bayesian model, but with some level of randomness in action selection related with the suboptimal behaviour commonly observed in mammals (see Morris et al., 2006 for behavioural results on suboptimal performance in monkeys).

We next analyse the DA responses to the two relevant stimuli sorting trials according to the definition of the task difficulty based on the Bayesian model. The mean firing rate in easy, intermediate and difficult classes reveals an interesting pattern of activation (see Figure 4.6). During the presentation of the base stimulus the firing rate in difficult classes is slightly lower than than in intermediate and easy classes, compatibly with the idea of a RPE coding. The effect is not significant during the entire period of the stimulus presentation but it does reach the significance level during the interval that lasts from 220 ms to 350 ms after the stimulus onset ($p < 0.05$; sliding ROC analysis with permutation test). During the presentation of the second stimulus the DA phasic activation clearly reflects the subjective difficulty that the animal needs to cope with in the decision process and the RPE coding related to this difficulty. In the temporal window from from 210 ms to 365 ms the firing rate in the three difficulty levels completely separate, resulting graduated according to the difficulty ($p < 0.05$; sliding ROC analysis with permutation

test). Indeed the mean activation during the entire 500 ms of stimulus presentation in easy classes is significantly higher that the activation in intermediate and difficult classes ($p < 0.05$,one-way ANOVA).

According to the Bayesian model, the pattern of responses observed during the presentation of the comparison frequency can be related the certainty of the animal about the future decision. The quantity that determine the choice according to the model, i.e the belief in Equation 4.1, can be indeed interpreted as the certainty of the animal about the decision $f_1 > f_2$. In difficult trials the belief about the decision is slightly above the threshold of 0.5 imposed by the MAP (maximum a posteriori) optimal criterion. This trials results therefore in a less accurate performance and in low levels of certainty that determine a less pronounced activation of DA neurons. We will further discuss the relationship between the activation of neurons and RPEs related to the certainty about the future decision in subsection 4.2.7.

4.2.6 DA response to the delivery of reward

In contrast to the response to the second stimulus, the response after the delivery of the reward does not depend on the difficulty of the task (see Figure 4.7a). This pattern of response is similar to that encountered in the detection task analysed in chapter 3. In that case the response to the reward delivery in hit trials did not depend on the amplitude of the relevant stimulus. However another recent study that analysed the response of DA neurons in the random dots motion (RDM) task showed an opposite result. In that case the response to a feedback tone announcing the trial outcome before the reward is delivered did depend on the motion coherence (Lak et al., 2017, Figure 3d right). Analysing the origin of these differences is beyond the scope of this work. However we speculate that the different pattern of responses could be related to the fact that in the RDM task the relevant stimulus is still present when the animal communicates its decision and receives the feedback tone. On the contrary both in the detection and in the discrimination task the relevant stimulus has already disappeared when the animal receives the reward. Therefore it is possible that in the first case (i.e. in the RDM tasks) the neural activity maintained some dependence on the certainty that determined the decision, whilst in the other to cases this dependence is lost.

In correct and error trials the DA firing rate behave as a RPE signal, resulting positive for correct trials and negative when the decision was wrong. It is interesting to note that the trend of activation separate depending on the trial outcome soon after the delivery of the reward. (Figure 4.7). This has to be contrasted with the response of DA neurons to

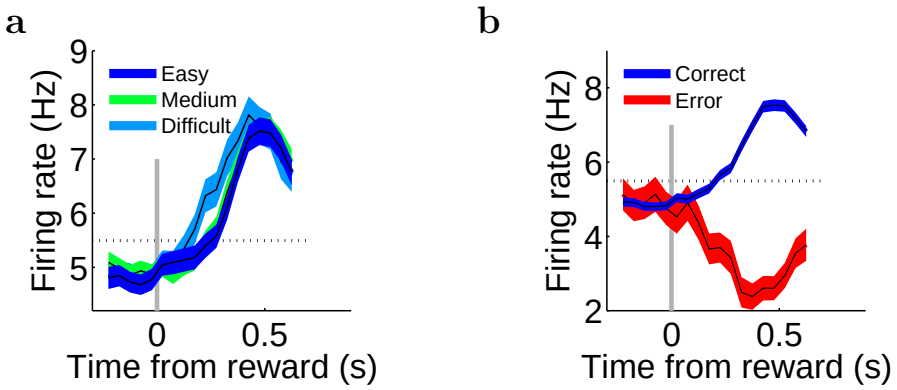


Figure 4.7: **DA response to the reward delivery.** (**a**) Mean population firing rate (black line, ± SEM colored band) plotted as a function of time for correct trials sorted according to the trial difficulty. Trials are aligned to the push button. The gray bar indicates the push button. (**b**) Same than in **a** but for correct and error trials.

the second stimulus; in that case the activity in both trial types followed a similar trend for about 150 ms.

4.2.7 Reinforcement learning model

We construct a computational model to investigate whether the processing of a Bayesian variable that, at behavioural level, is responsible for the contraction bias, could generate RPEs consistent with the the activity of DA neurons encountered in our data. As discussed in chapter 3 for tasks involving the elaboration of noisy sensory information, as the discrimination task analysed here, reinforcement learning (RL) models, and in particular the temporal difference (TD) algorithm, need to be based on partially observable Markov decision process (POMDP), as proposed in (Daw et al., 2006; Rao, 2010).

Similarly to what done in subsection 4.2.5 we assume that when presented with the first stimulus, the model uses the noisy observation o_1 to construct a belief over all the possible values of the first frequency, given by $\mathbf{b}_1(f) = P(f|o_1)$. To calculate the RPE at the onset of the first stimulus, the model also stores the values of waiting (W) until the presentation of the second stimulus given each possible value of f_1. These quantities are denotes as $Q(W, f_1)$. When the second stimulus is presented the model calculates a the belief over all possible values the second frequency, given by $\mathbf{b}_2(f) = P(f|o_2)$ using the noisy observation o_2. In addition it receives another noisy observation o_1^* about the frequency of the first stimulus, and construct a new belief over the value of the first frequency (see section 4.4

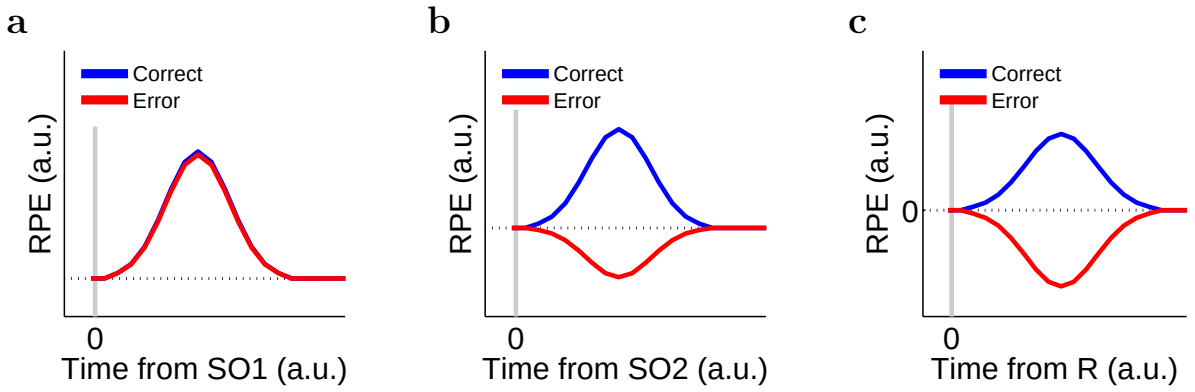


Figure 4.8: **Response of the model in correct and error trials.** (**a**) The RPE generated by the model after the onset of the first stimulus is similar in correct and error trials. (**b**) The RPE generated by the model after the onset of the second stimulus shows an activation in correct trials and a depression in error trials. (**c**) The RPE generated by the model after the reward delivery.

for more details about the model). We assume that the observation o_1^* is noisier than the observation o_2 to model the fact that the retrieval process is noisier than the sensory integration (as suggested in Ashourian and Loewenstein, 2011). Another set of variables is stored by the model after the presentation of the comparison stimulus: the values of calling the first frequency higher (H) or lower (L) given each possible pair of frequencies (f_1, f_2). These two set of variables are denoted as $Q(H, f_1, f_2)$ and $Q(L, f_1, f_2)$. In our model-based analysis we focus on the DA phasic activity, and therefore the RL model only produces three task-related prediction errors: the RPE at the onset of the first stimulus denoted as $\delta(f_1)$, the RPE at the onset of the first stimulus denoted as $\delta(f_2)$, and the RPE at the reward delivery denoted as $\delta(r)$. After the onset of the two stimuli the model combines Bayesian inference and information about the reward obtained in the previous trials (that is stored in the variables Q) to compute the RPEs. For the calculation of the RPE at the reward delivery we assume that the model does not use the information collected during the Bayesian inference (see section 4.4).

The RPE generated after the three relevant task events in correct and error trials is depicted in Figure 4.8. The RPE after the onset of the first stimulus is approximately independent from the final outcome (Figure 4.8a). However, similarly to the DA responses found in our data analysis (see Figure 4.5b), the RPE after the second stimulus results positive in correct trials and negative in error trials (Figure 4.8b). In the model this

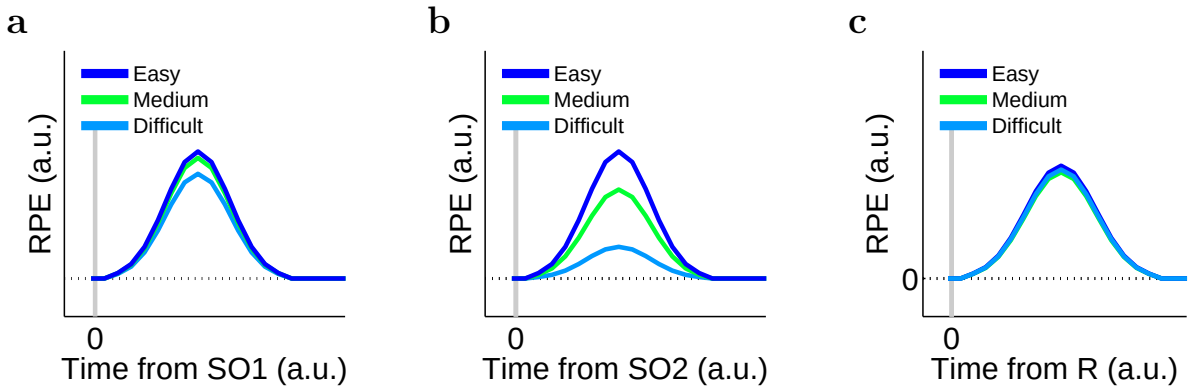


Figure 4.9: **Response of the model in correct trials sorted by difficulty.** (**a**) The RPE generated by the model after the onset of the first stimulus is similar in correct of different difficulty. (**b**) The RPE generated by the model after the onset of the second stimulus shows a graduation that depended on the subjective difficulty during the elaboration of the stimulus. (**c**) The RPE generated by the model after the reward delivery does not depend on the difficulty.

separation arises from the processing of the two Bayesian variables related to f_1 and to f_2. Before the onset of the second stimulus the reward expectation is associated with the first stimulus only. When the second stimulus is presented the reward expectation is associated with it becomes to be relevant. According to the model error trials produce because of ambiguous observation that generate two posterior distributions picked around similar values of frequencies and with great overlap. In such condition decisions are difficult and result in a great amount of incorrect choices. Therefore the onset of the second stimulus assumes low reward prediction values. The opposite is true for the majority of correct trial: the overlapping between the two posterior distribution is not much, the processing of the two belief results in a large amount of rewarded trials, and, as a consequence, the onset of the second stimulus assumes high reward prediction values. At the reward delivery the RPE shows a typical profile, resulting negative in error trials and positive in correct trials (Figure 4.8c).

Figure 4.9 depicts the RPE generated by the model in correct trials of different difficulty. The RPE after the onset of first stimulus show a slight dependence on the difficulty of the task. These RPEs are similar to the DA activations reported in Figure 4.6a. In the model, this profile produces because the information about the first frequency only is not enough to produce a fine prediction of future reward. However when the second

stimulus is presented the RPE clearly graduates according to the difficulty. This effect in the model is due to the different degree of overlapping between the two posterior distributions, similarly to what happens in correct and error trials. When the two distributions do not overlap at all (in easy trials), decision are always correct and the second stimulus generates great reward expectations and very pronounced RPEs. On the opposite when the overlapping is relevant (difficult trials), decision are sometimes wrong and the second stimulus generates lower reward expectations and little RPEs.

4.3 Discussion

Using the discrimination task, we found that the DA neurons responses showed a distinctive pattern of activation both at phasic and at a tonic level.

DA neurons were phasically activated by the base stimulus (Figure 4.3a). These phasic bursts did not encode the frequency of the stimulus, which represented the information that the animal needed to retain in working memory (Figure 4.2a). However we observed a slight tendency for a stronger activation for stimulus frequencies situated toward the center of the stimulus frequencies distribution. We attributed this effect to the contraction bias. According to this bias, the perceived frequency of the first stimulus shifts toward the center of its range, and, as a consequence, the frequencies that lie at the two extremes of the frequency range result in a worse performance. Indeed we found that the responses of DA neurons during the application of the first stimulus for central values of the frequency range, were (almost significantly) higher than those for extreme values. This pattern of activation suggested that the response to the first stimulus was coding a reward prediction error signal.

We next analysed the activation of neurons during the comparison stimulus. We focused our attention on the DA responses in correct and wrong trials. After an initial common increase, which lasted about 150 ms, the activity in wrong trials decreased significantly with respect to the activity in trials with correct choices (Figure 4.5b). A similar pattern of activity has been encountered in the RDM task (Lak et al., 2017). In that case, after the onset of the dots motions, DA response were uniform for about 200 ms, and then the mean firing rate increased for correct choices and decreased for error choices.

These responses to the second stimulus suggested that the DA activity could reflect some internal process related to the trial-to-trial difficulty faced by the animal. However, the variable controlling the trial difficulty is not clearly defined in the discrimination task, as confirmed by the different accurate performance in classes where the difference $|f_1 - f_2|$

was the same (see Figure 4.1b). We speculated that the effective difficulty in a given trial was not due to the real frequencies of the two stimuli, but to the perceived frequency of the first stimulus, that resulted shifted toward the center of the f_1 as a consequence of the contraction bias. We used a Bayesian model of the contraction bias (Ashourian and Loewenstein, 2011) to investigate the relationship between this behavioural anomaly and the activity of dopamine neuron. We adjusted the noise level of the model to get a good fit with the psychophysical results obtained in the experiment, and adopted the performance of the model as a measure of the difficulty of each frequencies pair. We found indeed that the response of neurons during the comparison stimulus was graduated according to the difficulty. It resulted higher for easy classes as posited by the reward prediction error coding commonly observed in the phasic DA activation.

We then tried better elucidate these responses by using a reinforcement learning (RL) model based on partially observable Markov decision processes (POMDPs). Guided from the result obtained at behavioural level the RL model used a posterior distribution (i.e a belief obtained using Bayesian inference) as input to calculate reward predictions and RPEs. The model reproduced all the relevant aspects of the phasic activation of the DA neurons. It suggested that the different responses to the second stimulus in correct and error trials could be due to a differential integration process, which changed from trial to trial. In correct trials of different difficulty the model explained the graduated responses to the comparison stimulus in terms of the belief about the higher frequency that the RL model used as input variable.

The fact that the two relevant stimuli were separated by a few seconds allowed to study the temporal profile of the DA activity during the temporal interval that require the use of working memory. We found a sustained positive tonic activation of the dopamine neurons during the delay period. This delay activation was not tuned to the first frequency value, and, unlike the response immediately after the base stimulus presentation, appeared unrelated with a reward prediction error coding. A previous study involving working memory did not report such a persistent activation during the delay period (Matsumoto and Takada, 2013). However, although we could not determine its origin, the positive modulation that we encountered is consistent with the idea that tonic dopamine plays an important role to retain the relevant information in working memory (Cohen et al., 2002). Moreover, this result is in line with persistent firing of dorsolateral prefrontal neurons that have long been implicated in working memory (Wilson et al., 1993; Rao et al., 1997).

Our results extended the understanding of the DA activity in tasks that require in two directions: on the one hand the positive modulation that we observed during the delay

period and that was not encountered in previous studies opens new perspective regarding the role of tonic DA firing in maintaining working memory functions in PFC. On the other hand the phasic responses to the relevant stimuli, apart from being in line with a reward prediction error coding, confirm that DA signals are triggered by internally arising experiences rather than external sensory stimulation per se (as suggested in chapter 3 and in some recent study; see for example Nomoto et al., 2010; de Lafuente and Romo, 2011; Matsumoto and Takada, 2013). In addition our analysis suggested that the DA responses to the comparison stimulus were shaped by the contraction bias.

4.4 Methods

Discrimination Task. Monkeys were trained to perform the vibrotactile task as depicted in section 4.4a . Trials began with the probe indenting the finger (Probe Down, PD), followed by the monkey grasping a metal bar with his other hand (Key Down,KD) to signal readiness. After a variable delay of 1500-3000 ms, the first stimulus was applied for 500 ms, followed by a 3000 ms delay and the second stimulus. The monkey then released the bar (Key Up,KU), indicated the discrimination by pressing one of two push-buttons with the right hand (Push Button, PB), and was rewarded for correctly discriminating the higher frequency. Stimuli were delivered to the skin of the distal segment of one digit of the restrained hand, via a computer-controlled stimulator (BME Systems; 2 mm round tip). Initial probe indentation was 500 μm. Vibrotactile stimuli were mechanical sinusoids. Stimulation amplitudes were adjusted to produce equal subjective intensities. Animals were handled in accordance with standards of the National Institutes of Health and Society for Neuroscience. All protocols were approved by the Institutional Animal Care and Use Committee of the Instituto de Fisiologa Celular.

Recordings. Recordings were obtained with quartz-coated platinum-tungsten microelectrodes (2 to 3 MΩ; Thomas Recording) inserted through a recording chamber located over the central sulcus, parallel to the midline. Midbrain DA neurons were identified on the basis of their characteristic regular and low tonic firing rates (1-10 spikes per second) and by their long extracellular spike potential (2.4 ms $\pm$ 0.4 SD). The group of 15 cells used for the analysis corresponded to those neurons that showed a phasic increase in discharge caused by the delivery of reward.

Data Analysis. For each neuron, we computed the firing rate as a function of time using 300 ms sliding windows displaced every. Responses to the first stimulus (in Figure 4.2) were measured in a 450 ms window centered 280 ms after the stimulus onset. The z-score in Figure 4.4a,b,d were measured during the entire delay period. All the responses (z-score) were standardized with respect to a temporal window preceding the onset of the base stimulus (the window lasted of 500 ms centered 1000 ms after the KD).

Bayesian model for the contraction bias in the discrimination task. In a Bayesian framework the discrimination task can be modeled as follows: when presented with a base stimulus of frequency f_1^* , the monkey has only access to a noisy representation of it, or observation o_1^*. The observation are sampled from a normal distribution with constant variance σ_1^2 around the true stimulus frequency; that is: if the true frequency of the first stimulus in a given trial is f_1^* the observation is sampled from $o_1^* = \mathcal{N}(f_1^*, \sigma_1)$. This noisy information of is combined with the knowledge of the prior distribution $P_1(f)$ to calculate the belief, or posterior distribution about the value of the first frequency $b_1(f) = P(f|o_1^*) \propto P_1(o_1^*|f) \cdot P_1(f)$. This is a vector of $n_1 = 6$ components (because the frequency of the first stimulus can assume 6 different values). The belief $b_2(f)$ about the second frequency (a vector of $n_2 = 12$ components) can be constructed in a similar way, with the only difference that the observation are now sampled from a normal distribution of variance σ_2^2. Given the two beliefs about the first and the second stimulus frequency, the posterior distribution about $f_1 > f_2$ can be obtained using the Equation 4.1. An ideal Bayesian observer, who has access to o_1 and o_2 , would report that $f_1 > f_2$ in trials in which $b(f_1 > f_2) > 0.5$. Therefore, the probability $P(H|f_1^*, f_2^*)$ that a model would report that f_1 was higher then f_2 in a trial in which f_1^* and f_2^* are presented is given by:

$$P(H|f_1^*, f_2^*) = \sum_{o_1} \sum_{o_2} P(o_1|f_1^*) P(o_2|f_2^*) G(b(f_1 > f_2)) \tag{4.2}$$

where $G(b(f_1 > f_2)) = 1$ if $b(f_1 > f_2) > 0.5$, and $G(b(f_1 > f_2)) = 0$ if $b(f_1 > f_2) < 0.5$. Equation 4.2 defines the performance of the model in any possible class. We adjust the two noise parameters σ_1 and σ_2 in order to minimize the difference between the performance of the model and the performance of the animal. To model the fact that the memory retrieval/decision making process and not during the memory encoding we constrained the two parameters to respect the relationship $\sigma_1 < \sigma_2$.

Reinforcement learning model. The model relies on the POMDP formalism. It assumes that a Bayesian module processes the sensory information about the values of

the two frequencies, and then transmits the result of this inference to a RL module that selects action and generates RPEs. In each trial at the onset of the first stimulus the Bayesian model receives an observation o_1 about the value of f_1, which is sample from a normal distribution of mean equal to the true frequency in that trial $\bar{f}_1$ and variance σ_S^2, i.e. $o_1 = \mathcal{N}(\bar{f}_1, \sigma_S)$. Using this observation it calculates the posterior over all possible value of f_1. For simplicity we assume an uniform prior. The belief distribution is a vector of n components, each one representing the probability that the presented frequency had specific value. Here we discretize the frequency range in steps of 1 Hz, so the dimension of the belief corresponds to the upper limit of the frequency range, that is 34 Hz for the first frequency. Given the belief distribution about f_1 the RL module calculates the value of waiting (W) until the presentation of the second stimulus:

$$Q(W|b_1(f)) = \sum_{f_1=1}^{34} Q(W, f_1) \cdot b_1(f_1) \tag{4.3}$$

The RPE at the onset of the first stimulus is calculated as $\delta(f_1) = Q(W|b_1(f))$.

At the onset of the second stimulus the Bayesian module calculates a posterior probability about the value of the second frequency $b_2(f)$, using the observation $o_2 = \mathcal{N}(\bar{f}_2, \sigma_S)$ (where $\bar{f}_2$ is the true value of f_2 in the trial). In addition it receives another observation $o_1^* = \mathcal{N}(\bar{f}_1, \sigma_R)$ from a storage area about the value of f_1. We assume $\sigma_R > \sigma_S$ to model the loss of information during the retrieval process. The observation o_1^* is used to calculate a new posterior about f_1, denoted as $b_1^*(f)$. These two beliefs are combined by the RL module to calculate the values of calling the first frequency higher (H) or lower (L) than the second one:

$$Q(H|b_1^*(f), b_2(f)) = \sum_{f_1=1}^{34} \sum_{f_2=1}^{44} Q(H, f_1, f_2) \cdot b_1^*(f_1) \cdot b_2(f_2) \tag{4.4}$$

$$Q(L|b_1^*(f), b_2(f)) = \sum_{f_1=1}^{34} \sum_{f_2=1}^{44} Q(L, f_1, f_2) \cdot b_1^*(f_1) \cdot b_2(f_2) \tag{4.5}$$

The outcome of the discrimination a is obtained as $a = \text{argmax}_D\, Q(D|b_1^*(f), b_2(f))$. This outcome is combined with the previous information about f_1 to compute the RPE a the onset of the second stimulus:

$$\delta(f_2) = Q(a|b_1^*(f), b_2(f)) - Q(W|b_1(f)) \tag{4.6}$$

The RPE in the above equation is used to update the set of values $Q(W, f_1)$ as follows:

$$Q(W, f_1) = Q(W, f_1) + \alpha \cdot \delta(f_2) \cdot b_1(f_1) \tag{4.7}$$

where α represents the learning rate.

We assume that between the offset of the second frequency and the second frequency and the delivery of the reward the system forgets the beliefs that determined the decision. Thus the RPE at the reward delivery using a decision value $Q(a)$ that is a weighted average of $Q(a|b_1^*(f), b_2(f))$ over uniform posterior distribution about the two stimuli:

$$\delta(R) = R - Q(a) \tag{4.8}$$

where $R = 1$ for correct discrimination and $R = 0$ otherwise. This RPE is is used to update the set of values $Q(a, f_1, f_2)$ as follows:

$$Q(a, f_1, f_2) = Q(a, f_1, f_2) + \alpha \cdot \delta(R) \cdot b_1^*(f_1) \cdot b_2(f_2) \tag{4.9}$$

Chapter 5

Final conclusions

The ability of processing and using incomplete information to predict the value of outcomes and make appropriate decisions is essential to the organization of behaviour. Although dopamine prediction errors are believed to play a crucial role in associative learning and goal-directed behaviour their role in decision making is yet to be clarified.

In this thesis we tried to elucidate how dopamine neurons behave during tasks that require non-trivial processing of external information. To do so, we combined the analysis of data from recordings of the firing activity of dopamine neurons taken while monkeys perform decision-making tasks using data analysis and modeling work based on algorithmic ideas from reinforcement learning. Importantly, we addressed this model-based analysis to two different experimental paradigms.

In chapter 3 we re-examined data recorded when the animal was engaged in the detection of possibly weak vibrotactile stimuli delivered at random times. This study allowed to shed light on several features of the dopamine reward prediction error signals which were not reported before. We demonstrated the existence of dopamine excitations to false detections of the vibrotactile stimulus, reward-related events the animal believed had occurred but which did not actually occurred. In line with a previous study (Carnevale et al., 2015) we found that the signature of these false detections in the dopamine activity became evident during the temporal window in which the stimulus was expected to happen. Building a Bayesian/reinforcement learning model we elucidated the way how the response of dopamine neurons were related to both timing and stimulus uncertainty. We showed that when the animal was asked to communicate its decision dopamine neurons coded a reward prediction error signal related to the certainty of the animal about the detection. In addition we found that the dopamine activity was shaped by temporal expectations in a way that depended on the subjective detection of the relevant stimulus.

In chapter 4 we analysed new recordings collected during a discrimination task in which the animal was presented with two vibrotactile stimuli and was required to distinguish the one with the higher frequency. We observed that dopamine neurons were phasically activated by the first stimulus. These phasic bursts did not code the specific information about the stimulus to be retained in working memory (the frequency), instead they were partially consistent with reward prediction error signals. The fact that the two relevant stimuli were separated by a few seconds allowed to study the temporal profile of the dopamine activity during the temporal interval that require the use of working memory. We found a sustained positive tonic activation of the dopamine neurons during the delay period. This delay activation was not tuned to the first frequency value, and, unlike the response immediately after the base stimulus presentation, appeared unrelated with a reward prediction error coding. A previous study involving working memory did not report such a persistent activation during the delay period (Matsumoto and Takada, 2013). However, although we could not determine its origin, the positive modulation that we encountered is consistent with the idea that tonic dopamine plays an important role to retain the relevant information in working memory (Cohen et al., 2002).

Tasks that require the sequential comparison of two stimuli separated by an interval of a few seconds constantly show a behavioural bias that is known as the contraction bias (first noticed in Hollingworth, 1910). Assuming that Bayesian computation underlies the contraction bias (Ashourian and Loewenstein, 2011) we found that the phasic activation of dopamine neurons after the second stimulus reflected this behavioural anomaly. Additionally, we used a Bayesian/reinforcement learning model to further elucidate the nature of this distinctive pattern of responses. The model results showed that the phasic bursts of dopamine neurons after the second stimulus can be interpreted as reward prediction error signals that stem from differential sensory evidence integration and that reflect the subjective difficult of the animal in processing the stimulus.

During the last years there has been an increasing interest in clarifying the role of dopamine in shaping behaviour. Many recent studies have focused on analysing the relationship between the dopamine activity and interval timing, showing that phasic and tonic dopamine signals conveyed information about the elapsed time (see for example Fiorillo et al., 2008; Bromberg-Martin et al., 2010; Pasquereau and Turner, 2015). Although apparently different from the decision making problems analysed in this thesis, from a theoretical perspective time is nothing more than any other state that cannot be observed directly (i.e a hidden state) that the system needs to infer in order to calculate reward predictions and reward prediction errors. Indeed, in line with this view, recent

data provided support for a temporal difference learning model that operates over belief states in a task that involved the estimation of time (Starkweather et al., 2017). Another recent study has demonstrated that, in addition to reflect interval timing, dopamine neuron activity can directly control the judgement of time (Soares et al., 2016). Therefore, regarding to the perception of time, the dopamine reward prediction error signals seem to operate on hidden states and actively guide the inference process.

In these thesis we analysed and modeled the dopamine signal in perceptual decision making tasks, mainly focusing on a form of state inference different from time inference. Our results support the view that also in complex decision making processes dopamine neurons operate on hidden states. The study suggests that dopamine neurons convey a teaching signal that reflects an optimal/Bayesian inference process and that is appropriate to guide optimal decision making. Although the way how these dopaminergic reward prediction error signals affect the inference process is yet to be clarify, our results open new perspectives on a possible active role of dopamine neurons in guiding optimal decision making.

Chapter 6

Conclusiones Finales

La habilidad de procesar y utilizar información incompleta para predecir el valor de dist- diferentes desenlaces y tomar las decisiones apropiadas es esencial para la organización del comportamiento. Aunque se piensa que los errores de predicción codificados por las neuronas dopamina juegan un papel crucial en aprendizaje asociativo y comportamiento orientado a objetivos, su rol en la toma de decisiones está aún por clarificar.

En esta tesis he intentado dilucidar como las neuronas dopamina se comportan en tareas que requieren procesamiento no trivial de información externa. Para hacer esto, he combinado análisis de datos sobre la tasa de disparo de neuronas dopamina, tomados en monos mientras estos completaban tareas de toma de decisiones, con trabajo de modelado inspirado en algoritmos de aprendizaje con refuerzo. He dirigido este análisis basado en modelos a dos paradigmas experimentales distintos.

En el capitulo 3, re-examinamo datos tomados mientras el animal esta involucrado en la detección de posibles estímulos débiles, administrados en tiempos aleatorios. Este estudio permitió descubrir varias características del comportamiento de las señales de predicción de error de la recompensa, que no se conocían con anterioridad. Demuestro la existencia de excitaciones de la dopamina a falsas detecciones del estimulo vibrotactil, eventos relacionados con recompensa que el animal creía que habían ocurrido, pero que no habían ocurrido realmente. En concordancia con un estudio previo (Carnevale et al,2015) encuentro que la traza en la actividad de la dopamina de estas falsos detecciones se vuelve notable durante la ventana de tiempo en la que se espera que ocurra el estimulo. Mediante un modelo que combina estimación bayesiana y aprendizaje con refuerzo he encontrado de que manera la respuesta de las neuronas dopamina esta relacionada con el cronometraje e incertidumbre del estimulo. Muestro que cuando se requería que el animal mostrase su decisión, la actividad de las neuronas dopamina reflejaba la certeza del animal respecto

a la detección. Asimismo encuentro que la actividad de la dopamina se ve afectada por las expectaciones temporales de una manera que depende de la detección subjetiva del estimulo relevante.

En el capítulo 4, he analizado nuevos datos tomados durante una tarea de discriminación, en la cual al animal se le presentan dos estímulos vibrotactiles y debe distinguir cual de los dos tiene mayor frecuencia. Observo que las neuronas se activan de forma fásica al primer estímulo. Estas ráfagas no codifican información especifica acerca del estímulo para ser retenida en la memoria de trabajo, sino que son parcialmente consistentes con errores de predicción de recompensa. El hecho de que los dos estímulos estuvieran separados por un intervalo de tiempo me ha permitido estudiar el perfil temporal de la actividad de la dopamina durante este intervalo, en el cual se requiere el uso de memoria de trabajo. Encuentro una activación positiva y sostenida de las neuronas dopamina durante el periodo entre estímulos. Esta actividad sostenida no depende de la frecuencia del primer estímulo, y a diferencia de la actividad justo después del primer estímulo, tampoco codifica errores de predicción de la recompensa. Estudios previos acerca de la memoria de trabajo no reflejaron esta actividad sostenida en el periodo entre estímulos (Matsumoto and Takada, 2013). No obstante, aunque no he conseguido encontrar su origen, esta modulación positiva de la actividad dopamina en el periodo entre estímulos es coherente con la idea de que la dopamina juega un papel importante en la memoria de trabajo (Cohen et al., 2002).

Aquellas tareas que requieren la comparación secuencial de dos estímulos separados por un periodo de unos pocos segundos presentan de manera consistente una tendencia en el comportamiento que se conoce como prejuicio de contracción (descrito por primera vez en Hollingworth, 1910). Asumiendo que lo que hay tras este prejuicio es integración bayesiana (Ashourian and Loewenstein, 2011), he encontrado que la activación fásica de las neuronas dopamina después del segundo estímulo reflejan esta anomalía en el comportamiento. Ademas, he utilizado un modelo que combina integración bayesiana y aprendizaje con refuerzo para entender la naturaleza de este distintivo patrón de respuesta. Los resultados del modelo nos mostraron que las ráfagas tónicas de la actividad de las neuronas dopamina después del segundo estimulo pueden interpretarse como señales que codifican errores en la predicción de recompensa, estas señales surgen de una integración diferencial de la evidencia sensorial y reflejan la dificultad subjetiva del animal en procesar el estímulo.

Durante los últimos años ha habido un interés creciente en dilucidar el papel que juega la dopamina en el comportamiento. Muchos estudios recientes se han centrado en analizar la relación entre actividad de la dopamina y la duración de los intervalos de la tarea,

mostrando que tanto la actividad fásica como tónica de la dopamina llevan información del tiempo transcurrido (ver por ejemplo Fiorillo et al., 2008; Bromberg-Martin et al., 2010; Pasquereau and Turner, 2015). Si bien parece un tema distinto a los analizados en esta tesis, el tiempo no es más que otro estado que no puede ser observado directamente (estado oculto), que el sistema necesita inferir para poder predecir recompensas futuras. En efecto, en concordancia con esta perspectiva, datos recientes proveen apoyo para un modelo de diferencia temporal que opera sobre estados de creencia en una tarea relacionada con la estimación del tiempo (Starkweather et al., 2017). Otro estudio reciente ha mostrado que ademas de reflejar el tiempo de los intervalos, las neuronas dopamina pueden afectar directamente la percepción del tiempo (Soares et al., 2016). De modo que, en relación a la percepción del tiempo, las señales de error en la predicción de la recompensa parecen operar sobre estados ocultos y guiar de manera activa el proceso de inferencia.

En esta tesis he analizado y modelado la señal de dopamina en actividades relacionadas con percepción y toma de decisiones, concentrándome principalmente en una forma de inferencia del estado diferente a la inferencia del tiempo. Mis resultados apoyan la idea de que también en tareas complejas de toma de decisiones las neuronas dopamina actúan en estados ocultos. Este estudio sugiere que las señales dopamina portan una señal de aprendizaje que refleja un proceso de inferencia bayesiana adecuado para guiar una óptima toma de decisiones. Aún cuando la manera en la cual estas señales de error en la predicción de la recompensa afectan el proceso de inferencia está aún por clarificar, miss resultados abren nuevas perspectivas en un posible rol activo de las neuronas dopamina en la toma de decisiones.

Appendix A

Supplemental Material to Chapter 2

Here I analyse the convergence of the TD algorithm for experiments of simple acquisition in which the reward follows a CS (assumed to occur at time 0) after a random ISI drawn from a probability distribution $f(t)$. The analysis of the convergence relies on the fact that the TD algorithm forces the mean of the delta signal across trials $\forall t$ to converge to zero.

The delta signal will always defined as (Montague et al., 1996; Ludvig et al., 2008):

$$\delta(t) = r(t) + \gamma V(t) - V(t-1) \tag{A.1}$$

I will assume that the ISI duration varies between t_r^{min} and t_r^{max}, i.e that $f(t) = 0 \forall t > t_r^{max}$. I will also use the implicit assumption of episodic tasks. This last condition forces the value function to zero at times bigger than t_r^{max}. I will denote with $h(t)$ the hazard of the reward at time t, i.e. the probability that the reward will occur at time t given that it has not yet occurred. The hazard can be expressed in term of the probability distribution $f(t)$ as $h(t) = f(t)/[1 - F(t)]$ where $F(t) = \sum_{i=0}^{t-1} f(i)$ is the cumulative distribution function of $f(t)$. Equivalently $h(t)$ can be written as $h(t) = f(t)/[\sum_{i=t}^{t_r^{max}} f(i)]$.

In the last section I will show that the TD model with the reset mechanism in the simulation of the simple acquisition described here produce a RPE at reward delivery that is equivalent (from an algorithmic point of view) to the RPE generated by the occurrence of a general task event that resets the representation of previous stimuli.

A.1 Convergence of the TD algorithm without reset

In absence of the reset mechanism δ signal in a given trial is in principle different from zero at each time step t. Considering that the reward is delivered at time t with probability $f(t)$, the constraint that the delta signal across trials $\forall t$ converges to zero implies the following equation:

$$< \delta(t) >= f(t) \cdot [1 - \gamma \cdot V(t) - V(t-1)] + [1 - f(t)] \cdot [0 + \gamma \cdot V(t) - V(t-1)] = 0 \quad \text{(A.2)}$$

where I use the fact that the scalar value of the reward is set to 1. Therefore the value function converges in mean to[1]:

$$V(t) = f(t+1) + \gamma \cdot V(t+1) \quad \text{(A.3)}$$

Given that $V(t_t^{max}) = 0$ the above equation implies that $V(t_r^{max} - 1) = f(t_r^{max})$ and that the value function can be written as:

$$V(t) = \sum_{i=0}^{N_t} \gamma^i f(t+1+i) \quad \text{(A.4)}$$

where I define $N_t = t_r^{max} - t - 1$. Equation A.4 corresponds to the expectation of the exponentially discounted sum of future rewards from time t to the end of the current trial. The above equation implies that when time elapses between t_r^{min} and t_r^{max} the value function decreases roughly as $[1 - F(t)]$. In particular when $\gamma = 1$ the value decreases exactly as $[1 - F(t)]$.

If time is represented by a tapped delay line (i.e through the CSC representation) the TD algorithm without reset converges exactly to the value function expressed in Equation A.4.

[1] Equation A.3 is equivalent to $V(t-1) = f(t) + \gamma \cdot V(t)$ that directly follows from Equation A.2.

A.2 Convergence of the TD algorithm with reset

The reset mechanism implies that when the reward is delivered at time t_r the value function $V(t) = 0$, and that the RPE $\delta(t+1) = 0$, $\forall t \geq t_r$. Therefore the δ signal in a given trial at time step t is different from zero if and only if the reward did not occur before such time step. This condition is verified with probability equal to the hazard $h(t)$ of the reward occurrence at time t. The constraint that the mean of the delta signal across trials $\forall t$ converges to zero implies that:

$$< \delta(t) >= h(t) \cdot [1 - \gamma \cdot 0 - V(t-1)] + [1 - h(t)] \cdot [0 + \gamma \cdot V(t) - V(t-1)] = 0 \quad \text{(A.5)}$$

The above equation takes into account that when the reward occurred at time t (this happens with probability $h(t)$ in each trial) $r(t) = 1$ and $V(t) = 0$, whereas when it does not occur $r(t) = 0$ and $V(t)$ can be different from zero.

Therefore the value function converges in mean to:

$$V(t) = h(t+1) + [1 - h(t+1)] \cdot \gamma \cdot V(t+1) \quad \text{(A.6)}$$

Considering that $V(t_r^{max}) = 0$ the recursive relationship in Equation A.6 can be written as:

$$V(t) = h(t+1) + \sum_{k=1}^{N_t} \gamma^k \cdot h(t+1+k) \prod_{i=1}^{k} [1 - h(t+1+k)] \quad \text{(A.7)}$$

where $N_t = t_r^{max} - t - 1$ as in section A.1. Using the fact that $h(t) = f(t) / \sum_{i=t}^{t_r^{max}} f(i)$ each term $h(t+k) \prod_{i=1}^{k} [1 - h(t+k-1)]$ of the sum in Equation A.7 can be written as $f(t+k) / \sum_{i=t}^{t_r^{max}} f(i)$, i.e as probability of the reward occurrence at time $t+k$ given that the reward has not occurred before t[2].

Defining $P(r(t+k)|r(t' < t) = 0) = f(t+k) / \sum_{i=t}^{t_r^{max}} f(i)$ the value function converges in mean to:

$$V(t) = \sum_{k=0}^{N_t} \gamma^k P(r(t+1+k) = 1 | r(t' < t) = 0) \quad \text{(A.8)}$$

It is easy to see that Equation A.8 is equivalent to Equation 2.7. When time is represented by a tapped delay line (i.e through the CSC representation) the TD algorithm with reset converges exactly to the value function expressed in Equation A.8 (see Figure A.1).

[2]Note that this probability is different from the hazard h(t+k) that represents the probability of the reward occurrence at time $t+k$ given that the reward has not occurred before $t+k$

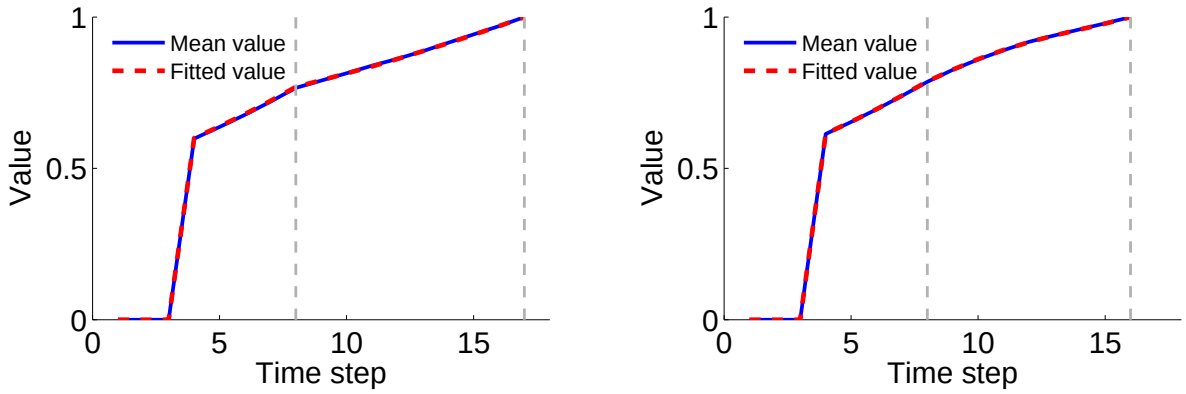


Figure A.1: Convergence of the value function with the reset mechanism. Time is represented through the CSC representation and reward is delivered following uniform distribution (Left) and a gaussian distribution(Right). The stimulus is delivered at $t_s = 3$ and the reward occurs between t_r^{min} and t_r^{max} (gray dashed lines). The blue line represents the mean value function obtained when simulating the TD model. The red dashed line represents the value function as obtained from Equation A.8.

A.3 TD learning and reset after a task event different from the reward

In this last section I will analyse the properties of the TD model with reset when the resetting task event is different from the reward. In order to allow analytic calculations I will adopt the simplified assumption that only the representation of one event (denoted as event j) is active when the resetting event i is presented.

I propose to study how the event i presentation affects the value developed by stimulus j, under the hypothesis that the event i resets the representation of the event j. Let assume that event j is presented at time $t_j^{on} = 0$ and that the event i follows the event j with probability $f_i(t)$ (different from 0 between t_i^{min} and t_i^{max}) and hazard denoted as $h_i(t)$. Denoting with $V_i(t)$ and $V_j(t)$ respectively the values of events i and j at time t the constraint that the delta signal across trials $\forall t$ converges to zero implies that:

$$< \delta(t) >= h_i(t) \cdot [\gamma V_i(t) - V_j(t-1)] + [1 - h_i(t)] \cdot [\gamma V_j(t) - V_j(t-1)] \tag{A.9}$$

where I use the fact that $r(t) = 0$, $V_i(t-1) = 0$, and that when the event i occurs $V_j(t) = 0$ (because of the reset).

Assuming that the TD model shows a stable behaviour after a sufficient number of trials the onset value of the event i converges to its asymptotic value denoted as V_i^* and this value is independent from the (variable) time elapsed before the occurrence of the event i. Equation A.9 the value function associated with the event j converges in mean to:

$$V_j(t) = \gamma V_i^* \cdot h_i(t+1) + [1 - h_i(t+1)] \cdot \gamma V_j(t+1) \tag{A.10}$$

Equation A.10 apart from the multiplicative factor γV_i^* is equivalent to Equation A.6. The value function $V_j(t)$ converges therefore in mean to:

$$V_j(t) = \gamma V_i^* \cdot \sum_{k=0}^{N_t^i} \gamma^k P(e_i(t+1+k) = 1 | e_i(t' < t) = 0) \tag{A.11}$$

where I Indicate as $P(e_i(t+k) = 1 | e_i(t' < t) = 0)$ the probability that the event i occurs at time $t + k$ given that it did not occur before time t, and $N_t^i = t_i^{max} - t - 1$.

Concerning to the dependence on temporal expectation Equation A.8 and Equation A.11 are perfectly equivalent, and therefore they are expected to show the same RPEs.

Appendix B

Supplemental Material to Chapter 3

B.1 Bayesian Module

Here we describe the detailed equations used by the bayesian module. This module represented some high-level cortical area receiving inputs from sensory areas. We referred to these inputs as observations x_t and interpreted them as Poisson trains with firing rates λ_i $(i = 0, ..., N_a)$. Each λ_i corresponded either to the absence of a vibrotactile stimulus (i = 0) or to the application of that stimulation with one of the $N_a = 9$ possible values of its amplitude during the time step t. Each of the 10 mean firing rates λ_i corresponded to a different state i of the world. In each time step t the module computed a posterior probability (belief) $b_t(i)$ about the hidden state of the world using the entire history of observations $X_{1:t}$ up to time t:

$$b_t(i) = P(\lambda_t = \lambda_i | X_{1:t}) \tag{B.1}$$

The beliefs about the absence and the presence of the stimulus corresponded respectively to:

$$\begin{aligned} b_t(sa) &= P(\lambda_t = \lambda_0 | X_{1:t}) \\ b_t(sp) &= \sum_{i \neq 0} P(\lambda_t = \lambda_i | X_{1:t}) \end{aligned} \tag{B.2}$$

Due to the complex temporal structure of the task evaluating the $b_t(i)$ required to estimate the joint posteriors $\tilde{b}_t(i, n)$ on the value of the rate of the input (λ_i corresponding to the state i) and the time n elapsed since the environment underwent a change to the state i. We therefore computed the belief over λ_t by marginalizing:

$$b_t(i) = \sum_n P(\lambda_t = \lambda_i, l_t = n | X_{1:t}) = \sum_n \tilde{b}_t(i, n) \tag{B.3}$$

We separated the last part of the history, i.e the last observation x_t, and calculated each belief $\tilde{b}_t(i,n)$ recursively over time using Bayes' rule:

$$\begin{aligned}\tilde{b}_t(i,n) &= P(\lambda_t = \lambda_i, l_t = n | X_{1:t-1}, x_t) \\ &= k \cdot P(x_t|\lambda_t = \lambda_i) \sum_n P(\lambda_t = \lambda_i, l_t = n | X_{1:t-1}) \end{aligned} \tag{B.4}$$

where $k = P(x_t|X_{1:t-1})$ is a normalization constant. The second term of Equation B.4 had been simplified using the Markov assumption and the fact that x_t did not depend on the length l_t (it depends only on the firing rate at current time λ_t). This term of Equation B.4 represented the observation probability (see section **??**) . The last term of the above equation could be rewritten as follow:

$$\begin{aligned}P(\lambda_t = \lambda_i, l_t = n | X_{1:t-1}) &= \sum_{j,m} \Big[P(\lambda_t = \lambda_i, l_t = n | \lambda_{t-1} = \lambda_j, l_{t-1} = m, X_{1:t-1}) \\ &\qquad P(\lambda_{t-1} = \lambda_j, l_{t-1} = m | X_{1:t-1}) \Big] \\ &= \sum_{j,m} \Big[P(\lambda_t = \lambda_i | \lambda_{t-1} = \lambda_j, l_{t-1} = m, l_t = n, X_{1:t-1}) \\ &\qquad P(l_t = n | \lambda_{t-1} = \lambda_j, l_{t-1} = m, X_{1:t-1}) \, \tilde{b}_{t-1}(j,m) \Big] \end{aligned} \tag{B.5}$$

Equation B.4 together with Equation B.5 represented a recursive relationship for the joint posteriors $\tilde{b}_t(i,n)$. Evaluating them required the knowledge of the change-point prior $CPP(l_t, l_{t-1}, \lambda_{t-1}, X_{1:t-1}, t-1) = P(l_t = n | \lambda_{t-1} = \lambda_j, l_{t-1} = m, X_{1:t-1})$ and of the transition probability $P(\lambda_t = \lambda_i | \lambda_{t-1} = \lambda_j, l_{t-1} = m, l_t = n, X_{1:t-1})$.

The change-point prior resulted independent from the history $X_{1:t}$ and, taking into account that the run-length either increased by one after each time step or became zero at a change point, the CPP could be defined as :

$$CPP(n, m, \lambda_j, t-1) = \begin{cases} 1 - h(\lambda_j, m, t-1) & \text{if} \quad n = m+1 \\ h(\lambda_j, m, t-1) & \text{if} \quad n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.6}$$

The function $h(\lambda_{t-1}, l_{t-1}, t-1)$ represented the hazard rate, i.e the probability that a change point occurred at time $t-1$ given that the state of the world was λ_{t-1} for exactly l_{t-1} time steps. It could be defined accordingly to the task structure (see Section **??**).

The third term of Equation B.5, i.e the transition probability, could be written as:

$$P(\lambda_t = \lambda_i | \lambda_{t-1} = \lambda_j, l_{t-1} = m, l_t = n) = \begin{cases} \delta_{ij} & \text{if} \quad n = m + 1 \\ T_{ij} & \text{if} \quad n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.7}$$

where we δ_{ij} represented the Kronecker delta and we introduced the matrix $T_{ij} = P(\lambda_t = \lambda_i | \lambda_{t-1} = \lambda_j, l_t = 0)$ representing the transition probability conditioned to the occurrence of a change-point (see section **??**).
Using Equation B.6 and Equation B.7 we could rewrite Equation B.4 as:

$$\begin{aligned} \tilde{b}_t(i, 0) &\propto \sum_{j \neq i} \sum_m T_{ij}\, h(\lambda_j, m, t-1) \tilde{b}_{t-1}(j, m) \\ \tilde{b}_t(i, n \neq 0) &\propto \big[1 - h(\lambda_i, n-1, t-1)\big] \tilde{b}_{t-1}(i, n-1) \end{aligned} \tag{B.8}$$

The equations above completely described the temporal evolution of the $\tilde{b}_t(i, n)$ once the hazard rate h and the transition probability matrix T_{ij} were defined.

B.1.1 Transition Probabilities

Given that the transition matrix T_{ij} was conditioned to the occurrence of a change-point we only needed to define the quantities $T_{i \neq 0, sa}$ and $T_{sa, i \neq 0}$. These probabilities were independent from the particular value of the firing rate λ_i in the stimulus present condition. We obtained that $T_{sp,sa} = 1/9$ (because all the 9 amplitude values were equally probable) and $T_{sa,sp} = 1$ (because the delay period always followed the stimulation).

B.1.2 Hazard Rate

As for the transition matrix the hazard rate for the stimulus present condition was independent from the particular value of the firing rate λ_i. The hazard rate only depended on the time $t - 1$, on the duration of an epoch before the transition l_{t-1} and on the state corresponding to that epoch λ_{t-1}

In stimulus absent condition this function took a value different from zero only during the possible stimulation windows and depended on and on the epoch length λ_{t-1} the time $t - 1$ (because transitions were not allowed during the delay period). We defined it as :

$$h(\lambda_{t-1} = \lambda_0, l_{t-1} = m, t-1) = \begin{cases} h_{sa}(m) & \text{if} \quad m = t - 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.9}$$

In stimulus present condition, given the task we were considering, the hazard rate only depended on the duration of the epoch before the transition and was defined simply as :

$$h(\lambda_{t-1} \neq \lambda_0, l_{t-1} = m, t-1) = h_{sp}(m) \tag{B.10}$$

The exact form of the functions $h_{sa}(m)$ and $h_{sp}(m)$ depended on the task temporal structure. If the interval timing mechanism was perfect the function $h_{sa}(m)$ would represent the hazard rate corresponding to a uniform probability density function whilst $h_{sp}(m)$ would represent the hazard rate corresponding to a fixed duration interval lasting as the stimulation period.

Nevertheless these definitions ignored the fact that animals' interval timing processes did not take place with infinite accuracy (the accuracy of temporal estimation is supposed to be constrained by Weber's law). Following (Janssen and Shadlen, 2005) we calculated a 'subjective' hazard function based on the assumption of timing scalar noise and used these subjective hazard to perform the inference. The value of the Weber fraction for time estimation used in the simulations was $\phi = 0.18$.

B.1.3 Observation Probabilities

The last step to implement Equation B.4 was to define the quantities $P(x_t|\lambda_t)$. We considered that the observation x_t represented the number of spikes produced in a sensory area on a given time step and it was generated from a poisson distribution with mean λ_t. The parameter λ represented the mean firing rate of a sensory area. Depending on the presence of the stimulus and on the amplitude value, the parameter λ_t could take the value λ_0, in stimulus absent conditions, and the value λ_i with $i \neq 0$, when a stimulus with amplitude i is presented. Therefore, we defined the observation x_t as follow:

$$x_t = \begin{cases} \text{Poisson}(\lambda_0) & \text{if the stimulus is absent} \\ \text{Poisson}(\lambda_i) & \text{if the stimulus is present with amplitude } i \end{cases}$$

We defined the probability to obtain the observation x_t given a mean firing rate λ_i at time t as:

$$\mathrm{P}(x_t|\lambda_i) = \mathrm{P}_{poisson}(x_t|\lambda_i) \tag{B.11}$$

where $\mathrm{P}_{poisson}(x|\lambda)$ indicated the probability to obtain the observation x given a poisson process with mean λ. The 10 values of the parameters λ_i were obtained from previously recorded data of the same experiment (de Lafuente and Romo, 2005) and corresponded to the mean firing rates of a sensory area in the 10 different conditions. Their values, ordered according to increasing values of the amplitude of the stimulus, were: 15 Hz, 15.2 Hz, 15.5 Hz, 16 Hz, 17 Hz, 20 Hz, 23 Hz, 27 Hz, 35 Hz and 40 Hz.

B.1.4 Belief Equations

Using Equation B.8 the posterior probability $b_t(i)$ of being in the state i could be expressed as:

$$\begin{aligned} b_t(i) &= \sum_n \tilde{b}_t(i,n) \\ &\propto \sum_{j\neq i} \sum_m T_{ij}\, h_j(m, t-1)\tilde{b}_{t-1}(j,m) \\ &+ \sum_{n\neq 0} \big[1 - h_i(n-1, t-1)\big]\tilde{b}_{t-1}(i, n-1) \end{aligned} \tag{B.12}$$

For the stimulus absent state the above equation took the form:

$$\begin{aligned} b_t(sa) &\propto \sum_{j\neq 0} \sum_m T_{sa,sp}\, h_j(m, t-1)\tilde{b}_{t-1}(j,m) \\ &+ \sum_{n\neq 0} \big[1 - h_{sa}(n-1, t-1)\big]\tilde{b}_{t-1}(sa, n-1) \end{aligned} \tag{B.13}$$

Using the fact that $\tilde{b}_t(sp,m) = \sum_{j\neq 0} \tilde{b}_t(j,m)$ and the considerations about the hazard rate and the transition probabilities made in the previous sections we obtained that:

$$\begin{aligned} b_t(sa) = k \cdot P(x_t|sa)\Big[&\sum_m h_{sp}(m)\tilde{b}_{t-1}(sp,m) + \sum_{n\neq t} \tilde{b}_{t-1}(sa, n-1) \\ &+ \big[1 - h_{sa}(l_{t-1} = t-1)\big]\tilde{b}_{t-1}(sa, t-1)\Big] \end{aligned} \tag{B.14}$$

The first two term of the Equation B.14 represented the probability of the delay interval whilst the last term corresponded to the probability of remaining within the pre-stimulus interval. Using Equation B.8 we could define $b_t(\lambda_i \neq \lambda_0)$ for each of the 9 amplitudes (with $\lambda_i \neq \lambda_0$) as follow:

$$\begin{aligned} b_t(i \neq 0) = k \cdot P(x_t|\lambda_i)\Big[&\sum_m T_{sp,sa} h_{sa}(t-1)\tilde{b}_{t-1}(sa, t-1) \\ &+ \sum_{n>0} \big[1 - h_{sp}(n-1)\big]\tilde{b}_{t-1}(i, n-1)\Big] \end{aligned} \tag{B.15}$$

Taking into account that $b_t(sp) = \sum_i b_t(i \neq 0)$ and the considerations about the transition probabilities and the hazard rate we obtained:

$$\begin{aligned} b_t(sp) = k \cdot \Big[1/9 \sum_i P(x_t|\lambda_i)\Big] &\sum_m h_{sa}(t-1)\tilde{b}_{t-1}(sa, t-1) \\ + k \cdot \Big[\sum_{n>0} \big[1 - h_{sp}(n-1)\big] &\sum_i P(x_t|\lambda_i)\tilde{b}_{t-1}(i, n-1)\Big] \end{aligned} \tag{B.16}$$

The former term in the above equation represented the probability of stimulus onset whilst the latter was the probability of remaining in a stimulus present state condition before the stimulus offset (but after the onset of the vibration).

The stimulus was detected by the bayesian module when the belief about its presence exceeded the belief about its absence:

$$b_t(sp) > b_t(sa) \implies \text{stimulus} \quad \text{detected} \tag{B.17}$$

B.2 Supplementary Figures

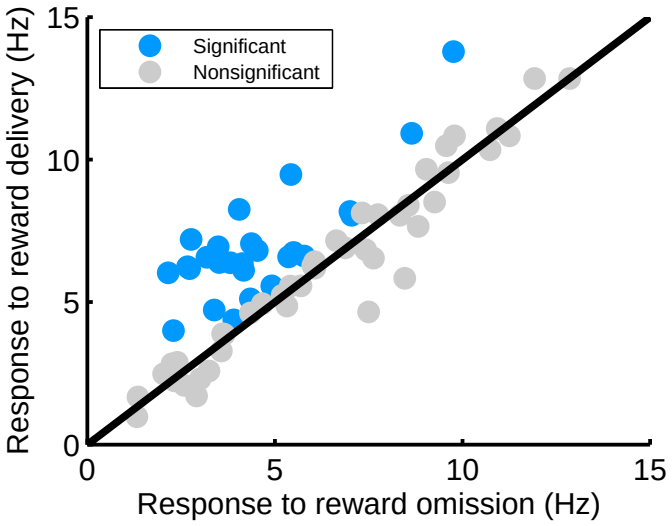


Figure B.1: **Selection of midbrain neurons.** The neurons used for the study (n=23) corresponded to those cells which responses to the reward delivery in correct trials were significantly higher of the responses to reward omission in incorrect trial ($P < 0.05$, two sample t test). Responses to the reward were measured in a 400 ms window centered 350 ms after the push button.

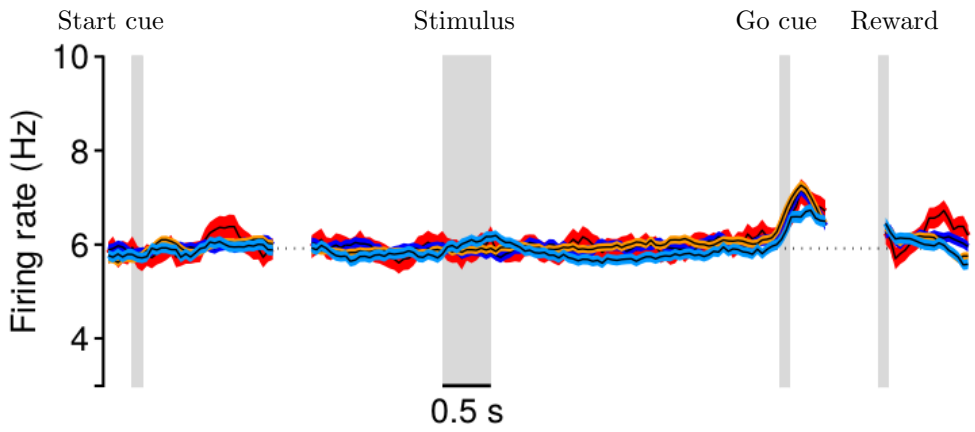


Figure B.2: **Mean firing rate of the discarded neurons.** Mean population firing rate (black line, ± SEM colored bands) of the discarded neurons plotted as a function of time for the four trial types. Activity is aligned to the start cue (left), go cue (center) and reward delivery (right). The dotted line indicates the baseline activity (5.9 spikes per second). The color code used to indicate the four trial types is the same as in Figure 3.1b.

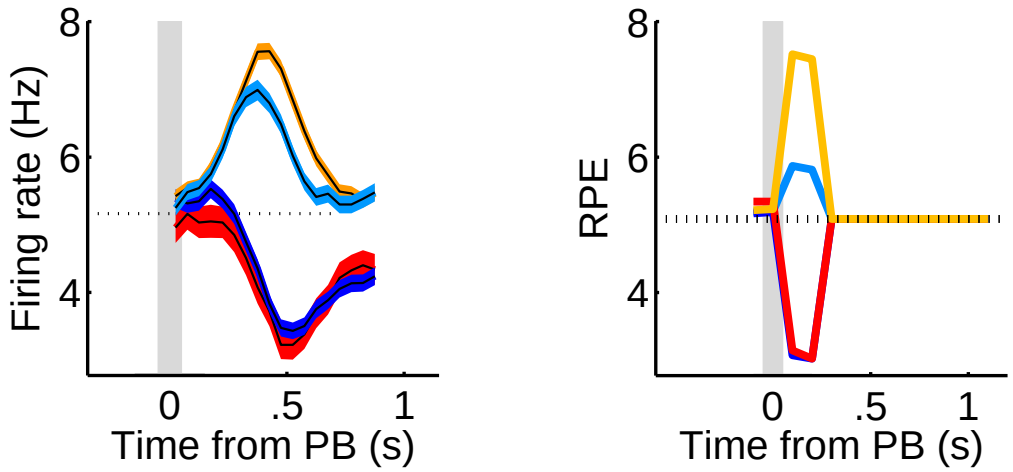


Figure B.3: **DA phasic responses and RPEs at the reward delivery.** Both the mean firing rate (left) and the RPE (right) showed a positive activation in rewarded trials and a pause in incorrect decision trials. The larger fraction of rewarded trials with stimulus-present decision was responsible for the smaller RPE in hit trials than in CR ones (right). The color code used to indicate the four trial types is the same as in Figure 3.1b. PB denotes the push button event.

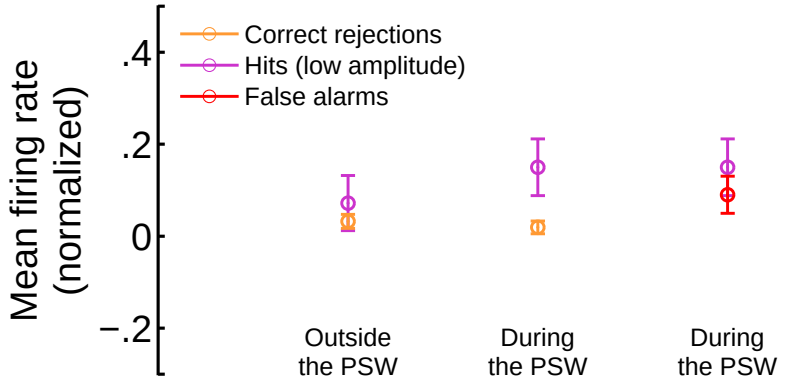


Figure B.4: **DA activity in low amplitude hit trials compared with the activity in stimulus absent trials.** The mean activity in low amplitude hit trials (see Methods) exhibited a significant positive modulation with respect to CR trials during the possible stimulation window (PSW) ($P < 0.05$, two sample one-tailed t test) but not outside it ($P = 0.26$, two sample one-tailed t test). Notably the activity in low amplitude hit trials and in FA trials during the PSW did not show any significant difference ($P = 0.21$, two sample one-tailed t test).

Bibliography

Abbott, L., DePasquale, B., and Memmesheimer, R.-M. (2016). Building functional networks of spiking model neurons. Nature neuroscience, 19, 350–355.

Adler, A., Katabi, S., Finkes, I., Israel, Z., Prut, Y., and Bergman, H. (2012). Temporal convergence of dynamic cell assemblies in the striato-pallidal network. The Journal of Neuroscience, 32, 2473–2484.

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annual review of neuroscience, 9, 357–381.

Alexander, W. H. and Brown, J. W. (2010). Hyperbolically discounted temporal difference learning. Neural computation, 22, 1511–1527.

Arnsten, A. F., Cai, J., Murphy, B., and Goldman-Rakic, P. (1994). Dopamine d 1 receptor mechanisms in the cognitive performance of young adult and aged monkeys. Psychopharmacology, 116, 143–151.

Ashourian, P. and Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. PloS one, 6, e19551.

Baddeley, A. D. and Hitch, G. (1974). Working memory. Psychology of learning and motivation, 8, 47–89.

Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. Journal of Neuroscience, 30, 9424–9430.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In Houk, James C and Davis, Joel L and Beiser, David G (Eds.), Models of information processing in the basal ganglia, pp. 215–232.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE transactions on systems, man, and cybernetics, pp. 834–846.

Barto, A. G., Sutton, R. S., and Watkins, C. J. (1989). Sequential decision probelms and neural networks. In NIPS, vol. 2, pp. 686–693.

Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron, 47, 129–141.

Bayer, H. M., Lau, B., and Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. Journal of Neurophysiology, 98, 1428–1439.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology, 191, 391–431.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). Neuro-Dynamic Programming. (Athena Scientific).

Bogacz, R. and Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. Neural computation, 23, 817–851.

Brody, C. D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cerebral cortex, 13, 1196–1207.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. Neuron, 67, 144–155.

Brown, J., Bullock, D., and Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. Journal of Neuroscience, 19, 10502–10511.

Cai, X., Kim, S., and Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron, 69, 170–182.

Calabresi, P., Gubellini, P., Centonze, D., Picconi, B., Bernardi, G., Chergui, K., Svenningsson, P., Fienberg, A. A., and Greengard, P. (2000). Dopamine and camp-regulated

phosphoprotein 32 kda controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. Journal of Neuroscience, 20, 8443–8451.

Carnevale, F., de Lafuente, V., Romo, R., Barak, O., and Parga, N. (2015). Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty. Neuron, 86, 1067–1077.

Carnevale, F., de Lafuente, V., Romo, R., and Parga, N. (2012). Internal signal correlates neural populations and biases perceptual decision reports. Proceedings of the National Academy of Sciences, 109, 18938–18943.

Carnevale, F., de Lafuente, V., Romo, R., and Parga, N. (2013). An optimal decision population code that accounts for correlated variability unambiguously predicts a subject's choice. Neuron, 80, 1532–1543.

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nature neuroscience, 19, 111–116.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., and Haynes, J.-D. (2017). The distributed nature of working memory. Trends in Cognitive Sciences.

Cohen, J. D., Braver, T. S., and Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. Current opinion in neurobiology, 12, 223–229.

Cook, E. P. and Maunsell, J. H. (2002). Dynamics of neuronal responses in macaque mt and vip during motion detection. Nature neuroscience, 5, 985–994.

Daw, N. D. (2003). Reinforcement learning models of the dopamine system and their behavioral implications. Ph.D. thesis, Citeseer.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. Neural computation, 18, 1637–1677.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience, 8, 1704–1711.

Daw, N. D. and Touretzky, D. S. (2000). Behavioral considerations suggest an average reward td model of the dopamine system. Neurocomputing, 32, 679–684.

Dayan, P. (1992). The convergence of td (λ) for general λ. Machine learning, 8, 341–362.

Dayan, P. and Abbott, L. F. (2001). Theoretical neuroscience, vol. 10. (Cambridge, MA: MIT Press).

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience, 8, 429–453.

de Lafuente, V. and Romo, R. (2005). Neuronal correlates of subjective sensory experience. Nature neuroscience, 8, 1698–1703.

de Lafuente, V. and Romo, R. (2006). Neural correlate of subjective sensory experience gradually builds up across cortical areas. Proceedings of the National Academy of Sciences, 103, 14266–14271.

de Lafuente, V. and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proceedings of the National Academy of Sciences, 108, 19767–19771.

de Lafuente, V. and Romo, R. (2012). Dopaminergic activity coincides with stimulus detection by the frontal lobe. Neuroscience, 218, 181–184.

Ding, L. and Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. Neuron, 79, 640–649.

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. Nature neuroscience, 11, 966–973.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science, 299, 1898–1902.

Friedrich, J. and Lengyel, M. (2016). Goal-directed decision making with spiking neurons. The Journal of Neuroscience, 36, 1529–1546.

Friedrich, J., Urbanczik, R., and Senn, W. (2011). Spatio-temporal credit assignment in neuronal population learning. PLoS Comput Biol, 7, e1002092.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E., and Bestmann, S. (2012). Dopamine, affordance and active inference. PLoS Comput Biol, 8, e1002327.

Gershman, S. J., Moustafa, A. A., and Ludvig, E. A. (2014). Time representation in reinforcement learning models of the basal ganglia.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences, 108, 15647–15654.

Hanes, D. P. and Schall, J. D. (1996). Neural control of voluntary movement initiation. Science, 274, 427.

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. Psychological Bulletin, 97, 35.

Hernández, A., Nácher, V., Luna, R., Zainos, A., Lemus, L., Alvarez, M., Vázquez, Y., Camarillo, L., and Romo, R. (2010). Decoding a perceptual decision process across cortex. Neuron, 66, 300–314.

Hernández, A., Zainos, A., and Romo, R. (2002). Temporal evolution of a decision-making process in medial premotor cortex. Neuron, 33, 959–972.

Hinton, S. C. and Meck, W. H. (2004). Frontal–striatal circuitry activated by human peak-interval timing in the supra-seconds range. Cognitive Brain Research, 21, 171–182.

Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nature Neuroscience, 1, 304–309.

Hollingworth, H. L. (1910). The central tendency of judgment. The Journal of Philosophy, Psychology and Scientific Methods, 7, 461–469.

Houk, J., Adams, J., and Barto, A. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement, models of information processing in the basal ganglia (eds. jc houk, jl davis and dg beiser), 249/270.

Huang, Y. and Rao, R. (2013). Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. PLoS ONE, 8.

Janssen, P. and Shadlen, M. N. (2005). A representation of the hazard rate of elapsed time in macaque area lip. Nature neuroscience, 8, 234–241.

Jin, D. Z., Fujii, N., and Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. Proceedings of the National Academy of Sciences, 106, 19156–19161.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. Neural networks, 15, 535–547.

Joel, D. and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience, 96, 451–474.

Jun, J. K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C. D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. Journal of Neuroscience, 30, 916–929.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial intelligence, 101, 99–134.

Knill, D. C. and Richards, W. (1996). Perception as Bayesian inference. (Cambridge University Press).

Kobayashi, S. and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. The Journal of Neuroscience, 28, 7837–7846.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., and Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Current Biology, 27, 821–832.

Lau, B. and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. Neuron, 58, 451–463.

Lee, A. M., Tai, L.-H., Zador, A., and Wilbrecht, L. (2015). Between the primate and 'reptilian'brain: rodent models demonstrate the role of corticostriatal circuits in decision making. Neuroscience, 296, 66–74.

Lee, D., Seo, H., and Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. Annual review of neuroscience, 35, 287–308.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of neurophysiology, 67, 145–163.

Ludvig, E. A., Bellemare, M. G., and Pearson, K. G. (2011). A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives. Computational neuroscience for advancing artificial intelligence: Models, methods and applications, pp. 111–144.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural computation, 20, 3034–3054.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. Cognitive, Affective, & Behavioral Neuroscience, 9, 343–364.

Matsumoto, M. and Takada, M. (2013). Distinct representations of cognitive and motivational signals in midbrain dopamine neurons. Neuron, 79, 1011–1024.

Meck, W. H. (2006). Neuroanatomical localization of an internal clock: a functional link between mesolimbic, nigrostriatal, and mesocortical dopaminergic systems. Brain research, 1109, 93–107.

Mello, G. B., Soares, S., and Paton, J. J. (2015). A scalable population code for time in the striatum. Current Biology, 25, 1113–1122.

Merchant, H., Harrington, D. L., and Meck, W. H. (2013). Neural basis of the perception and estimation of time. Annual review of neuroscience, 36, 313–336.

Mirenowicz, J. and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. Journal of neurophysiology, 72, 1024–1027.

Moghaddam, B., Adams, B., Verma, A., and Daly, D. (1997). Activation of glutamatergic neurotransmission by ketamine: a novel step in the pathway from nmda receptor blockade to dopaminergic and cognitive disruptions associated with the prefrontal cortex. Journal of Neuroscience, 17, 2921–2927.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. The Journal of neuroscience, 16, 1936–1947.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron, 43, 133–143.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nature Neuroscience, 9, 1057–1063.

Murphy, B., Arnsten, A., Goldman-Rakic, P., and Roth, R. (1996). Increased dopamine turnover in the prefrontal cortex impairs spatial working memory performance in rats and monkeys. Proceedings of the National Academy of Sciences, 93, 1325–1329.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. Neuron, 41, 269–280.

Niv, Y. (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology, 53, 139–154.

Niv, Y., Duff, M. O., and Dayan, P. (2005). Dopamine, uncertainty and td learning. Behavioral and Brain Functions, 1, 6.

Niv, Y. and Montague, P. R. (2008). Theoretical and empirical studies of learning. Neuroeconomics: Decision making and the brain, pp. 329–50.

Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. The Journal of Neuroscience, 30, 10692–10702.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. science, 304, 452–454.

Packard, M. G. and Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. Annual review of neuroscience, 25, 563–593.

Pasquereau, B. and Turner, R. S. (2015). Dopamine neurons encode errors in predicting movement trigger occurrence. Journal of neurophysiology, 113, 1110–1123.

Potjans, W., Morrison, A., and Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. Neural Computation, 21, 301–339.

Pouget, A., Drugowitsch, J., and Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. Nature neuroscience, 19, 366–374.

Rao, R. P. (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. Frontiers in computational neuroscience, 4, 146.

Rao, R. P., Olshausen, B. A., and Lewicki, M. S. (2002). Probabilistic models of the brain: Perception and neural function. (MIT press).

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. Science, 276, 821–824.

Reynolds, J. N. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. Neural Networks, 15, 507–521.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience, 10, 1615–1624.

Roitman, J. D. and Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of neuroscience, 22, 9475–9489.

Romo, R., Brody, C. D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature, 399, 470–473.

Romo, R., Hernández, A., Zainos, A., Lemus, L., and Brody, C. D. (2002). Neuronal correlates of decision-making in secondary somatosensory cortex. Nature neuroscience, 5, 1217–1225.

Romo, R. and Salinas, E. (2003). Flutter discrimination: neural codes, perception, memory and decision making. Nature Reviews Neuroscience, 4, 203–218.

Romo, R. and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. Journal of neurophysiology, 63, 592–606.

Rossi-Pool, R., Salinas, E., Zainos, A., Alvarez, M., Vergara, J., Parga, N., and Romo, R. (2016). Emergence of an abstract categorical code enabling the discrimination of temporally structured tactile stimuli. submitted.

Sawaguchi, T. (2001). The effects of dopamine and its antagonists on directional delay-period activity of prefrontal neurons in monkeys during an oculomotor delayed-response task. Neuroscience research, 41, 115–128.

Sawaguchi, T. and Goldman-Rakic, P. S. (1991). D1 dopamine receptors in prefrontal cortex: involvement in working memory. Science, 251, 947–951.

Schönberg, T., Daw, N. D., Joel, D., and O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. Journal of Neuroscience, 27, 12860–12867.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of neurophysiology, 80, 1–27.

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. Physiological reviews, 95, 853.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. The Journal of Neuroscience, 13, 900–913.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. Journal of neuroscience, 12, 4595–4610.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 1593–1599.

Schultz, W. and Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. Journal of neurophysiology, 63, 607–624.

Setlow, B., Schoenbaum, G., and Gallagher, M. (2003). Neural encoding in ventral striatum during olfactory discrimination learning. Neuron, 38, 625–636.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: seeing and deciding. Proceedings of the national academy of sciences, 93, 628–633.

Shadlen, M. N. and Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. Journal of neurophysiology, 86, 1916–1936.

Shankar, K. H. (2015). Generic construction of scale-invariantly coarse grained memory. In Australasian Conference on Artificial Life and Computational Intelligence, pp. 175–184. Springer.

Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of time. Neural Computation, 24, 134–193.

Shankar, K. H. and Howard, M. W. (2013). Optimally fuzzy temporal memory. Journal of Machine Learning Research, 14, 3785–3812.

Soares, S., Atallah, B. V., and Paton, J. J. (2016). Midbrain dopamine neurons control judgment of time. Science, 354, 1273–1277.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society of London B: Biological Sciences, 265, 2015–2020.

Starkweather, C. K., Babayan, B. M., Uchida, N., and Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience, 20, 581–589.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nature neuroscience, 16, 966–973.

Suri, R. E. (2002). Td models of reward predictive responses in dopamine neurons. Neural networks, 15, 523–533.

Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Experimental brain research, 121, 350–354.

Suri, R. E. and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience, 91, 871–890.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine learning, 3, 9–44.

Sutton, R. S. and Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement.

Sutton, R. S. and Barto, A. G. (1998). Reinforcement learning: An introduction, vol. 1. (MIT press Cambridge).

Takahashi, Y. K., Langdon, A. J., Niv, Y., and Schoenbaum, G. (2016). Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat vta depends on ventral striatum. Neuron.

Tank, D. and Hopfield, J. (1987). Neural computation by concentrating information in time. Proceedings of the National Academy of Sciences, 84, 1896–1900.

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. Science, 307, 1642–1645.

Todd, M. T., Niv, Y., and Cohen, J. D. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In Advances in neural information processing systems, pp. 1689–1696.

Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. IEEE transactions on automatic control, 42, 674–690.

Tsitsiklis, J. N. and Van Roy, B. (1999). Average cost temporal-difference learning. Automatica, 35, 1799–1808.

Tsitsiklis, J. N. and Van Roy, B. (2002). On average versus discounted reward temporal-difference learning. Machine Learning, 49, 179–191.

van Seijen, H. and Sutton, R. S. (2014). True online td (lambda). In ICML, vol. 14, pp. 692–700.

Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., and Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. PLoS Comput Biol, 5, e1000586.

Vergara, J., Rivera, N., Rossi-Pool, R., and Romo, R. (2016). A neural parametric code for storing information of more than one sensory modality in working memory. Neuron, 89, 54–62.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. Nature, 412, 43–48.

Wencil, E. B., Coslett, H. B., Aguirre, G. K., and Chatterjee, A. (2010). Carving the clock at its component joints: neural bases for interval timing. Journal of neurophysiology, 104, 160–168.

Wickens, J., Begg, A., and Arbuthnott, G. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. Neuroscience, 70, 1–5.

Wickens, J. R., Budd, C. S., Hyland, B. I., and Arbuthnott, G. W. (2007). Striatal contributions to reward and decision making. Annals of the New York Academy of Sciences, 1104, 192–212.

Williams, G. V. and Goldman-Rakic, P. S. (1995). Modulation of memory fields by dopamine d1 receptors in prefrontal cortex. Nature, 376, 572.

Wilson, F. A., Scalaidhe, S. P., and Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. SCIENCE-NEW YORK THEN WASHINGTON-, 260, 1955–1955.